frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 05 December 2025
DOI 10.3389/frai.2025.1648073

OPEN ACCESS

EDITED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

REVIEWED BY
Madis Jürviste,
Institute of the Estonian Language, Estonia
Yogi M. Kom,
Lancang Kuning University, Indonesia

*CORRESPONDENCE
Hadi Khalilia
✉ hadi.khalilia@unitn.it;
✉ h.khalilia@ptuk.edu.ps

RECEIVED 16 June 2025
REVISED 31 October 2025
ACCEPTED 13 November 2025
PUBLISHED 05 December 2025

CITATION
Khalilia H, Otterbacher J, Bella G, Darma S and Giunchiglia F (2025) Crowdsourcing lexical diversity. *Front. Artif. Intell.* 8:1648073.
doi: 10.3389/frai.2025.1648073

Crowdsourcing lexical diversity

Hadi Khalilia[1,2]*, Jahna Otterbacher[3], Gábor Bella[4],
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]CYENS - Centre of Excellence, Open University of Cyprus, Nicosia, Cyprus, [4]IMT Atlantique,
Lab-STICC UMR CNRS 6285, Brest, France

Lexical-semantic resources (LSRs), such as online lexicons and wordnets, are fundamental to natural language processing applications as well as to fields such as linguistic anthropology and language preservation. In many languages, however, such resources suffer from quality issues: incorrect entries, incompleteness, but also the rarely addressed issue of bias toward the English language and Anglo-Saxon culture. Such bias manifests itself in the absence of concepts specific to the language or culture at hand, the presence of foreign (Anglo-Saxon) concepts, as well as in the lack of an explicit indication of untranslatability, also known as cross-lingual *lexical gaps*, when a term has no equivalent in another language. This paper proposes a novel crowdsourcing methodology for reducing bias in LSRs. Crowd workers compare lexemes from two languages, focusing on domains rich in lexical diversity, such as kinship or food. Our LingoGap crowdsourcing platform facilitates comparisons through microtasks identifying equivalent terms, language-specific terms, and lexical gaps across languages. We validated our method by applying it to two case studies focused on food-related terminology: (1) English and Arabic, and (2) Standard Indonesian and Banjarese. These experiments identified 2,140 lexical gaps in the first case study and 951 in the second. The success of these experiments confirmed the usability of our method and tool for future large-scale lexicon enrichment tasks.

KEYWORDS

multilingual lexicon, language diversity, crowdsourcing, linguistic gap, lexical typology

1 Introduction

Despite advances in deep learning and large language models, lexical-semantic resources (LSRs)—such as WordNet (Miller, 1995)—remain essential for natural language processing (NLP) tasks, including machine translation, word sense disambiguation, and information retrieval (Katsuta and Yamamoto, 2020; Loureiro and Jorge, 2019; Barbouch et al., 2021), as well as for supporting broader research domains such as linguistic anthropology, cultural linguistics, and language documentation.

Due to its dominance, the English language and, in particular, Princeton WordNet (PWN) (Miller, 1995), has played a distinguished role in the construction of lexical databases for many languages. The English lexicon has been widely adopted as a *pivot* representation of lexical meaning across languages, but also as the source language for *translation-based* lexicon development (Bond and Foster, 2013). Relying on PWN as a "standard" meaning inventory, however, leads to the creation of resources that suffer from *language modeling bias*, due to deep-running linguistic and cultural differences across speaker communities (Giunchiglia et al., 2017, 2023; Bella et al., 2024). For instance, English and Italian lack an equivalent equivalent for the Arabic word خالة, which means "*mother's sister*," whereas Arabic lacks a term for *nephew* (expressed in Italian as *nipotino*) (Khalilia et al., 2023). Such instances of *lexical diversity*, referred to as *(cross-lingual) lexical gaps* by Giunchiglia et al. (2018), occur when a word in one language lacks a counterpart in another. When English is used as the reference language, language-specific concepts and lexical gaps may remain undocumented. Yet, cross-lingual NLP applications must account for phenomena of linguistic diversity (Giunchiglia et al., 2017). For example, machine translation systems often encounter lexical gaps. Google Translate and ChatGPT mistakenly render "*do not give cider to your child*" into Arabic as لا تعطي عصير التفاح لطفلك, which means "*do not give apple juice to your child*," reflecting Arabic's lack of a term for *cider*. This highlights the challenge of achieving lexical equivalence across languages.

Addressing lexical diversity requires a *systematic* approach to building diversity-aware datasets. To our knowledge, our expert-driven approach (Khalilia et al., 2023) remains the only method that enables lexical gap identification at an advanced level within the target language, particularly in contexts where experts possess domain-specific knowledge. However, a major limitation of this approach is its *unidirectional* design (English → Target Language), which reinforces an *English bias* and overlooks culture-specific concepts in non-English languages. Additionally, the reliance on professional linguistic experts significantly limits its applicability to *low-resource languages*, thereby restricting coverage of global linguistic diversity.

Crowdsourcing has emerged as an effective means for developing NLP and linguistic resources, particularly those reflecting general language usage by native speakers. Prior efforts have included parallel corpus construction (Post et al., 2012), query systems like CrowdDB (Franklin et al., 2011), WordNet development (Ganbold et al., 2018), lexicon enhancement (Nair, 2022), word sense disambiguation (Parent and Eskenazi, 2010), sentiment analysis (Kasumba and Neumman, 2024), and information retrieval (Lease and Yilmaz, 2012). In this paper, we aim to provide two key contributions:

1. A *novel crowdsourcing methodology* for exploring lexical diversity across language pairs within specific semantic domains [e.g., food (Ashley et al., 2004), kinship (Khishigsuren et al., 2022), and body parts (Wierzbicka, 2007)]. The method involves: (a) semi-automated generation of lexical entries for each language, (b) crowdsourcing evaluations by native speakers who compare lexical entries to identify meaning equivalents and lexical gaps, and (c) validation by ordinary native speakers, followed by expert verification.
2. *Empirical validation* of our method via two case studies involving English–Arabic and Indonesian–Banjarese language pairs, focused on food-related terminology. Across 132 tasks with 36 workers, we identified 2,140 lexical gaps in English–Arabic (1,532 in Arabic, 608 in English) and 951 (750 in Banjarese, 201 in Indonesian) in Indonesian–Banjarese, along with 1,957 equivalent terms.

Our methodology is innovative in four key aspects:

1. *Language independence*: It applies to *any language pair*, regardless of existing linguistic resources (e.g., lexical databases, encyclopedias, or digital and undigitized dictionaries or corpora).
2. *No reliance on pivot languages*: It does not depend on *English* or any other language as an *intermediary*.
3. *Bidirectional exploration*: It supports *comparative* analysis from both source to target and vice versa.
4. *Applicability to both human and machine agents:* It can be implemented using either native-speaking crowd workers or large language models (LLMs). Our experiments show that native speakers are more effective than LLMs in identifying culturally and linguistically specific concepts, particularly in low-resource language contexts.

The structure of the paper is as follows. In Section 2, we review previous research related to our study. Section 3 presents an overview of lexical diversity and lexical gaps. Our crowdsourcing methodology is described in Section 4, followed by its implementation and evaluation in Section 5. This includes the introduction of the LingoGap platform in Section 5.1, two case studies on food-related terminology in English–Arabic and Indonesian–Banjarese (Section 5.2 and 5.3), and a comparison of crowdsourced data quality with LLM-generated annotations in Section 5.4. In Section 6, we discuss the use of crowdsourcing for constructing diversity-aware datasets. Finally, we conclude the paper in Section 7.

2 Related work

Crowdsourcing has been widely employed to create various linguistic resources, including lexical-semantic data. For instance, Ganbold et al. (2018) developed a Mongolian WordNet using a two-phase crowdsourcing workflow via the CrowdCrafting[1] platform. In the translation phase, volunteers suggested synonymous words by translating English PWN synsets into Mongolian. The subsequent validation phase employed inter-rater agreement metrics, such as Fleiss' kappa and Krippendorff's alpha, to ensure quality, achieving a precision of 0.74 for 947 synsets. Similarly, Wijesiri et al. (2014) bootstrapped a Sinhala WordNet from English with the help of bilingual internet users. Lanser et al. (2016) created a Japanese lexicon from DBpedia by first constructing an English version and then using annotators on CrowdFlower for translation.

Benjamin and Radetzky (2014) introduced a mobile app–based crowdsourcing model "*Fidget Widget*" to develop lexicons for low-resource languages. Biemann and Nygaard (2010) used Amazon Mechanical Turk (MTurk) to collect word senses for building a sense inventory from scratch. El-Haj et al. (2015) recruited annotators to construct the Essex Arabic Summaries Corpus, yielding 2,360 sentences and 41,493 words in Jordanian and Gulf Arabic.

Other efforts include (Manerkar et al., 2022), who developed "*Konkani Shabdarth*," a crowdsourcing platform allowing community members (e.g., students and faculty members) to enhance to the Konkani WordNet by adding missing words to its synsets. Fišer et al. (2014) introduced SloWCrowd to correct errors in the Slovene WordNet (Gantar and Krek, 2011), while Čibej and Arhar Holdt (2019) used crowdsourcing via PyBossa[2] to clean the Thesaurus of Modern Slovene. Nair (2022) proposed a Google Forms–based

[1] https://crowdcrafting.org/

[2] https://pybossa.com/

mobile approach for enhancing the Malayalam WordNet, referencing PWN.

The conventional method of *expanding* WordNet through translation (Fellbaum and Vossen, 2012) often fails to capture culture-specific concepts. For example, Arabic WordNet (Freihat et al., 2024) translates *uncle* as عم "*father's brother*," omitting maternal uncles. Bahasa WordNet (Noor et al., 2011) maps *sister* to *kakak* "*elder sibling*," reducing semantic precision. In contrast, MultiWordNet (Pianta et al., 2002), which employs a *merge* strategy with bilingual dictionaries, explicitly captures lexical gaps but lacks coverage in rich semantic domains like kinship and food, and has since been discontinued.

Lexical typology, a subfield of linguistics, investigates cross-linguistic diversity by examining how languages encode meaning within specific semantic domains (Plungyan, 2011). Lexical-typological research has explored translation-related challenges, particularly the presence or absence of lexicalized concepts across languages. Prior studies have focused on semantic domains known for considerable cross-linguistic variation, such as kinship terminology (Kemp and Regier, 2012), color categories (Roberson et al., 2005), food-related terms (Ashley et al., 2004), human body parts (Wierzbicka, 2007), and actions like cutting and breaking events (Majid et al., 2007) or putting and taking (Kopecka and Narasimhan, 2012). Despite ongoing research, publicly available datasets in this area remain scarce. Notable exceptions include Murdock (1970)'s kinship classification, which has been incorporated into D-PLACE (Kirby et al., 2016), and aspects of Kay and Cook (2016)'s research on color terminology, available in the lexicon section of the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). Another example is a dataset on color categorization by McCarthy et al. (2019), accessible via GitHub.[3]

Digital lexicons are increasingly employed in lexical typology, enabling researchers to analyze a broader range of languages and semantic domains. A notable example is the KinDiv[4] lexicon (Khishigsuren et al., 2022), which includes 1,911 lexical items and documents 37,370 lexical gaps related to kinship across 699 languages. Our study (Khalilia et al., 2023) builds on this resource, specifically examining kinship-related lexical diversity in Arabic dialects and Indonesian languages. Other relevant research includes Åke Viberg (1984)'s foundational study on perceptual vocabulary across 50 languages, which was later expanded by Georgakopoulos et al. (2022) to incorporate data from 1,220 languages.

Only one previous attempt to crowdsource lexical gaps is known—Giunchiglia et al. (2015), who developed a platform for translating English lexicalizations into Italian, including identifying lexical gaps, with the help of *linguistic experts*. In contrast, our approach leverages *non-expert native speakers* and differs significantly in its design and objectives.

Key differences between our approach and prior work include:

1. *No reliance on English as an intermediary:* Unlike methods that use English as a *pivot*, our methodology compares datasets from any two languages directly, thereby avoiding *English-centric bias*.
2. *Focus on lexical diversity and bidirectional exploration:* We exclusively target *lexical diversity* and conduct *bidirectional* comparisons to identify lexical gaps without assuming a fixed source-target direction. Contributions from native speakers are collected via micro-tasks enabled by our crowdsourcing platform.

3 Lexical diversity and cross-lingual lexical gaps

Translation is a complex process influenced by cultural and lexical diversity, often resulting in challenges when striving for meaning equivalence across languages (Catford, 1965; Bella et al., 2022). Vocabulary embedded in specific cultural contexts—such as the Arabic terms for meals during Ramadan, a month in which Muslims fast from sunrise to sunset—demonstrates these challenges. For instance, السحور *suhur*, "*a pre-dawn meal consumed before the daily fast begins*," and الافطار *iftar* "*the meal eaten at sunset to break the fast*," encapsulate cultural practices that have no direct equivalents in many other languages. Similarly, culturally bound terms related to alcohol in European languages—such as the English *Bitter*, "*a dry, sharp-tasting ale with a strong flavor of hops*," the Italian *Amaretto*, "*an almond liqueur*," and the German *Weizenbock* "*a wheat beer of bock strength*." In Indonesia, rice is not merely a food item but a staple central to daily life and national identity. Indonesians use various terms to describe its forms: *Gabah* "*harvested but unhulled rice*," *Beras* "*uncooked rice*," *Nasi* "*cooked rice*," *Kerak Nasi* "*scorched or crispy rice stuck to the bottom of the pot*." These examples illustrate how languages encode distinct worldviews and culturally specific concepts that often resist direct translation.

In this study, we examine *cross-lingual lexical gaps*—a phenomenon in which a word in the source language lacks a direct and precise equivalent in the target language. Such gaps often arise from cultural or regional specificities unique to individual linguistic communities and typically resist systematic translation through established rules or patterns (Lehrer, 1970). A lexical gap is formally defined as follows:

Definition 1 (Lexical Gap). Let L_1 and L_2 be two natural languages, and let $w \in L_1$ be a lexical item expressing a well-defined meaning m. A lexical gap from L_1 to L_2 exists if there is no lexical item $w' \in L_2$ such that w' conveys m without semantic loss, approximation, or periphrasis.

To illustrate this, Table 1 presents examples of lexical gaps in food-related concepts across five languages. As the table shows, no single language offers concise lexicalizations for all the listed concepts, yet each concept is lexicalized in at least one language. These cross-linguistic variations pose challenges for both human and machine translation. Furthermore, substituting culturally specific terms with more general or approximate equivalents may lead to unintended meanings.

For example, the Arabic word الأسودان "*water and dates*" has no direct equivalent in English, as illustrated in our experiment in Section 5.2. This lexical gap can lead to mistranslations, such as

[3] https://github.com/aryamccarthy/basic-color-terms

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

TABLE 1 Lexicalizations of four meanings around the concept of (food) in five languages.

Meaning	English	Japanese	Arabic	Italian	Indonesian
savory taste	GAP	うま味	GAP	GAP	gurih
firm pasta texture	GAP	GAP	GAP	al dente	GAP
water and dates	GAP	GAP	الأسودان	GAP	GAP
crispy	crispy	GAP	GAP	croccante	renyah

The gray shading marks cells where the meaning is not lexicalized—creating a lexical gap—in that language. In other words, the concept listed in the first column exists as a language-independent meaning (idea), but the language does not have an equivalent word for it; therefore, it is represented in the table as "GAP."

Google Translate's rendering[5] of the Arabic sentence يتناول الأسودين يبدأ الصائم إفطاره as "*The fasting person begins his breakfast by eating lions.*" Another instance involves a mistranslation by ChatGPT-4 of the Indonesian word *Kembili*—"*a root vegetable similar to a potato*"—into Banjarese as *Umbi-Umbian*, "*a broader term referring to various tuberous vegetables.*" Such inaccuracies highlight the need of identifying and addressing lexical gaps to preserve semantic integrity in cross-linguistic communication.

Recognizing patterns of lexical diversity highlights the need for scalable methods to document lexical gaps across a broad range of languages and semantic domains. However, existing approaches often depend on expert-driven processes that are *English-centric* and *unidirectional*—as discussed in Section 2—which limits their scalability and accessibility, particularly for low-resource languages. To address these limitations, the following section introduces *a novel crowdsourcing methodology* that leverages native speaker insights to systematically identify and verify lexical gaps.

4 Crowdsourcing methodology

This section outlines a methodology for the crowd-based collection of evidence related to lexical diversity. Lexical diversity is inherently an interlingual phenomenon, referring to meanings or distinctions that are not shared across languages. Such phenomena are especially common—but not limited to—socially or culturally significant domains, including food, religion, family relationships (Khishigsuren et al., 2022), motion verbs (Wälchli and Cysouw, 2012), body parts (Wierzbicka, 2007), colors (McCarthy et al., 2019), spatial dimensions (Lang, 2001), cutting and breaking events (Majid et al., 2007), pain predicates (Reznikova et al., 2012), perception verbs (Åke Viberg, 1984), and putting and taking events (Kopecka and Narasimhan, 2012). The *initial inputs* to our methodology are, on the one hand, an *ordered source–target language pair*, and on the other hand, a *semantic field* (SF). The language pair is ordered because the methodology yields direction-dependent results: for the pair (A, B), it identifies lexicalizations present in A but absent in B, and for (B, A), the reverse—lexicalizations present in B but not in A. The *output* of the method is a list of words in the source language (SL), each word annotated as either a lexical gap or lexicalized in the target language (TL), with an equivalent word provided in the latter case. This methodology defines two key roles: the task requester and the crowd worker. The *task requester* is responsible for several critical functions: (1) constructing the input datasets for both the SL and TL within the SF, (2) designing and creating the crowdsourcing tasks, (3) overseeing task execution to ensure high-quality contributions, and (4) validating and exporting the finalized crowdsourced data. In contrast, the *crowd worker* contributes by identifying equivalent terms and lexical gaps in the TL. The methodology is structured into three main steps:

1. *Task generation*: A semi-automated process that produces two lists of *lexical entries*, one for each language. A lexical entry is a tuple (*word*, *gloss*)—that is, a term paired with its definition.
2. *Crowdsourcing*: A *crowdsourcing micro-task* consists of one lexical entry from the SL, along with all lexical entries from the TL as candidates. The goal is to determine whether the SL entry is translatable to the TL and, if so, to provide the equivalent TL entry. Our policy requires that crowd workers be native speakers of the TL and possess sufficient command of the SL. By "sufficient command," we mean the ability to accurately understand the SL lexical entry, potentially with the aid of lexicons or online resources. In contrast, deciding whether an equivalent exists in the TL is a more challenging task that demands deep familiarity and active knowledge of the language—hence the requirement for TL-native workers.
3. *Validation*: This step consolidates the contributions of crowd workers through a native speaker validation process, followed by expert-based verification.

4.1 Step 1: task generation

Task generation takes lexical resources and corpora as input and produces a list of micro-tasks (i.e., *word–gloss* tuples) for the workers. This process is semi-automatic: candidate micro-tasks are generated algorithmically and then filtered by an expert. The primary challenge lies in identifying candidate words that belong to the target semantic field in a manner that is robust across a wide range of languages and dialects, including low-resourced ones. We achieve this robustness by leveraging a variety of language resources, depending on what is available for the given language and semantic field. These resources may include mono- or multilingual lexical databases (e.g., wordnets), online dictionaries and encyclopedias (e.g., Wikipedia, Wiktionary, traditional dictionaries), language models or word embeddings, digital or undigitized corpora, and fieldwork with native speakers.

Lexical databases, such as the UKC[6] or other wordnets, are preferred data sources due to their rich, structured, and meaning-annotated entries with definitions. Features such as domain tags

[5] Translation retrieved on February 17, 2025.

[6] https://ukc.datascientia.eu

and hierarchical relations support the filtering of entries by semantic field. However, only a few languages possess large-scale, high-quality lexical databases, and even fewer offer systematically annotated glosses.

Online dictionaries and encyclopedias serve two purposes: (1) providing glosses when they are absent from lexical databases—by extracting the first sentence from Wiktionary or Wikipedia for a given word, or discarding the word if no entry is found (in which case, no task is generated); and (2) offering candidate words potentially related to the semantic field, which are later filtered using vector similarity methods.

Language models and word embeddings are used to generate lexical entries when no adequate lexical resource exists in the source language for the given semantic field. A word list—obtained from a dictionary, as described above—is filtered to identify terms belonging to a specific semantic field (e.g., *cake*, *bread*, *tomato* for *food*). A vector representation of the field—constructed from a definition and example terms drawn from a small corpus—is generated using word or sentence embeddings (e.g., AraBERT for Arabic). These vectors are then compared to dictionary word vectors using cosine similarity, and the most similar words are selected.

Text corpora can be used to train word embeddings or language models when pre-trained versions are not available for a given language. For languages without existing digital corpora, an initial corpus digitization step is required. Finally, *fieldwork*, though often overlooked, is an effective method for obtaining a high-quality, focused corpus of words and definitions for a given language. It is particularly useful for low-resource languages and dialects when direct contact with native speakers is possible. The process begins with a language expert providing an initial set of seed words belonging to the semantic field. Native speakers then contribute additional words and definitions related to the field and the seed words.

For widely spoken languages such as English, Spanish, or Chinese, task generation is possible using a lexical DB alone (e.g., the UKC), although any combination of the aforementioned resources and methods may also be applied. For languages where lexical DBs are of lower quality, offer inadequate coverage, or do not exist—such as Arabic or Hungarian—data from lexical DBs can be supplemented with entries from traditional lexicons, filtered using language models. For even lower-resourced languages that lack language models but have usable corpora—such as European minority languages and dialects—digital dictionaries can provide candidate input words, and word embeddings can be trained and used to filter them according to the semantic field. Finally, for dialects and severely endangered languages with few or no existing corpora, results from prior fieldwork—such as kinship terms for Arabic dialects (Khalilia et al., 2023)—can be used to produce smaller-scale but potentially high-quality word lists.

During the data preparation phase, SL terms are first collected from predefined semantic fields (e.g., Food, Emotion). For each SL entry, definitions are automatically extracted from multiple lexical resources such as Wiktionary, Wikipedia, and the UKC. When definitions are unavailable in these resources, alternative methods—such as corpus-based extraction, linguistic databases, or fieldwork—are employed to ensure comprehensive lexical coverage. As noted in Lines 233–235, *any combination of the aforementioned resources and methods may be applied* to generate semantic field words and their corresponding definitions.

Once the SL definitions are established, TL candidate terms are retrieved semi-automatically from lexical databases using the same procedure applied to the SL terms. SL–TL pairs are then formed into tuples that serve as input for the subsequent crowdsourcing phase. Each tuple includes the SL word, its definition, and a list of TL candidates with their respective definitions.

For instance, within the Food semantic field, the English source term *banana* is extracted from the SL lexical list together with its definition retrieved from Wiktionary. The initial TL candidates—such as موز (Arabic)—are then automatically generated using an electronic dictionary. These preliminary SL–TL tuples are subsequently presented to annotators through the LingoGap interface, where participants validate the suggested equivalents, refine them as needed, or provide new TL terms if no appropriate match is found.

4.2 Step 2: crowdsourcing

This section describes how the requester engages crowd workers to identify lexical gaps and equivalent terms between the SL and the TL. Crowd workers, recruited via a selected crowdsourcing platform, utilize the datasets created in the previous step. Since lexical diversity can arise in both the SL and the TL, the crowdsourcing process is conducted twice—once in each direction. In the first experiment, SL lexical entries—comprising word–gloss tuples—are mapped to the TL. In this phase, crowd workers identify equivalent terms as well as lexical gaps in the TL. The second experiment reverses the direction: the TL from the first experiment is treated as the new SL, and the original SL becomes the TL. Previously identified equivalents (i.e., overlapping terms) are excluded, and the focus shifts to mapping the remaining TL entries to the SL. Task crowdsourcing is organized into three phases, as described below:

1. *Crowd selection*: The requester selects proficient crowd workers to participate in the task.
2. *Contribution collection*: Selected crowd workers complete micro-tasks on a crowdsourcing platform to provide equivalent terms and identify lexical gaps by comparing SL entries with TL entries, and vice versa.
3. *Contribution quality control*: The requester applies real-time quality control mechanisms to ensure the reliability and consistency of the collected data.

4.2.1 Crowd selection

Ensuring high-quality responses from crowd workers is essential. We adopt a two-step selection process[7]:

(1) *Proficiency Test:* A preliminary test (comprising 10%–25% of the total questions) evaluates worker capability on the same platform as the main task (Liu et al., 2013). To ensure domain-specific reliability, the test questions are contextualized within

[7] Platforms like Prolific and MTurk offer built-in tools for quality control (Robinson et al., 2019).

the semantic domain of the crowdsourcing task. Specifically, they assess workers' understanding of domain-relevant vocabulary and concepts. For example, when the task involves the food domain, the test includes food-related terms; for a kinship domain, it focuses on kinship terminology. This domain-oriented design ensures that selected workers demonstrate both linguistic proficiency and domain-specific competence—or conceptual familiarity—necessary for producing high-quality annotations.

(2) *Filtering via Krippendorff's Alpha:* To identify low-quality annotators, we compute inter-annotator agreement (IAA) using Alpha,[8] which accounts for chance agreement, multiple annotators, and missing data. A subset of questions—contextualized within the food-related domain and comprising 25% of the total items—is annotated by proficient workers and a linguistic expert to apply this filtering method (see Figure 1).

The worker filtering process is conducted on LingoGap—a custom-built crowdsourcing platform designed to collect lexical diversity data from non-expert native speakers—with input from a linguistic expert. The process begins by identifying the SL, TL, and SF, followed by the selection of approximately 10 SL terms within the chosen SF. After establishing an IAA threshold, workers are invited to complete a task on LingoGap. Workers whose agreement score (Alpha) meets or exceeds the threshold are classified as high-quality. For those who fall below the threshold, a subset of workers is evaluated—together with the expert—using all possible combinations of contributors (ranging from individual workers to all combinations except one), using the mathematical combination function described in Brualdi (2004). Subsets that meet the IAA threshold are retained as high-quality, while excluded workers are categorized as low-quality.

Example

Consider a group of three workers, $G_1 = \{w_1, w_2, w_3\}$, performing a task. If their agreement (measured by Alpha) exceeds a predefined threshold, all are considered high-quality.

If not, an expert (*Exp*) helps identify the low-quality worker. The task is repeated with the expert and each pair of workers:

$$\{Exp, w_1, w_2\}, \quad \{Exp, w_1, w_3\}, \quad \{Exp, w_2, w_3\}$$

If any of these combinations meets the threshold, the excluded worker is flagged as low-quality. If none meet the threshold, the task is repeated with the expert and each individual worker:

$$\{Exp, w_1\}, \quad \{Exp, w_2\}, \quad \{Exp, w_3\}$$

If one combination passes, the other two workers are marked low-quality. If none pass, all workers in G_1 are classified as low-quality.

4.2.2 Contribution collection

In this step, we use *LingoGap*, described in Section 5.1, a crowdsourcing tool developed to identify lexical gaps and equivalent words in a given language pair.

A requester—using the *admin* interface—creates a task and configures its details, including the description, language pair, and date. He or she selects source lemmas and their glosses from the SF dataset constructed in the source language. Additionally, He or she provides comprehensive *instructions* and clear *guidelines* for crowd workers through a customized spreadsheet template (see Table 2), which includes nine default guidelines. These guidelines can be added, edited, or removed depending on the language pair involved in the experiment. For instance, one guideline prohibits the use of machine translation for defining words.

Once a task is created, crowd workers access the *worker* interface[9] and follow the provided guidelines to answer multi-step questions for each source word, presented sequentially. Each word prompts three *multiple-choice questions (MCQs)*. The utility of MCQs for domain-targeted tasks has been demonstrated by Welbl et al. (2017) in lexical semantic evaluations. The example below illustrates a semantic equivalence task for the English word "*cider*" in comparison with Arabic:

Question: "*Does the Arabic language include an equivalent meaning to the English word described below? If yes, please write the Arabic word along with its definition.*"
Word: "*cider*" **Definition:** "*a beverage made from juice pressed from apples.*"

- Choice 1: Yes, word: سايدر, definition: مشروب كحولي مصنوع من التفاح
- Choice 2: No
- Choice 3: Don't know

Crowd workers can select one of three options: (1) an equivalent meaning—Choice 1, retrieved from a precompiled Arabic food lexicon created during task setup; (2) a lexical gap—Choice 2; or (3) uncertainty—Choice 3, "Don't know." Since the concept of "*cider*" does not exist in Arabic, the correct answer is a lexical gap.

4.2.3 Contribution quality control

To ensure the reliability of crowdsourced data, we implement two live quality control mechanisms:

(1) *Attention check questions (ACQs).* ACQs are embedded within regular tasks to assess worker attentiveness and compliance with instructions. These simple questions are designed to detect careless responses. Following Liu et al. (2013), we include one ACQ for every ten questions and require workers to achieve at least 90% accuracy, consistent with the threshold recommended by Robinson et al. (2019).

(2) *Completion time monitoring.* The time taken to complete a question serves as an indicator of engagement and response quality. Extremely short or long durations may suggest inattentiveness or rushed work. LingoGap logs completion times and automatically filters out outliers that deviate significantly from a worker's average, ensuring only reliable data are retained.

[8] Throughout this paper, the term "Alpha" refers specifically to Krippendorff's Alpha.

[9] http://lingogap.disi.unitn.it/

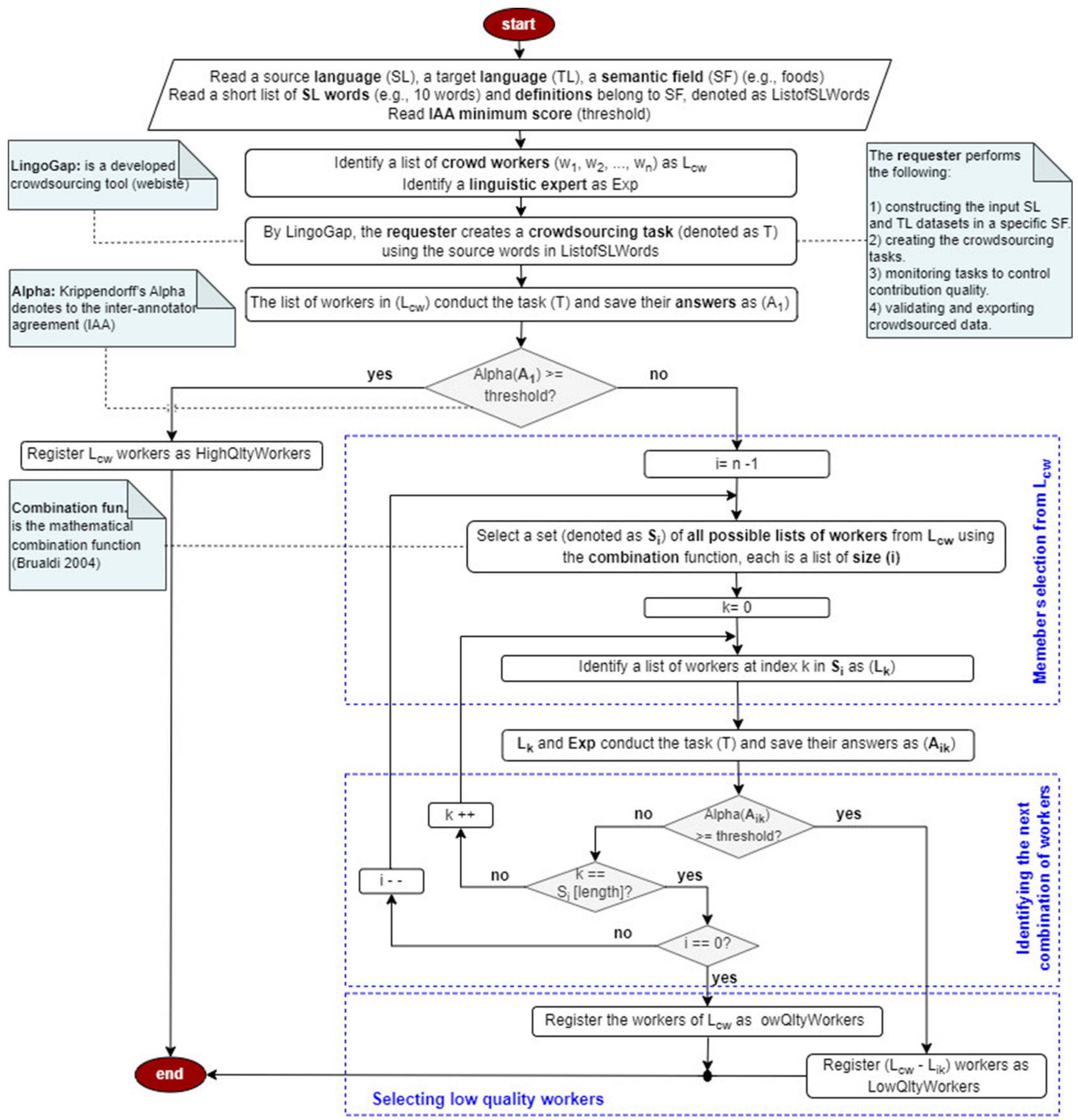

start
Read a source language (SL), a target language (TL), a semantic field (SF) (e.g., foods)
Read a short list of SL words (e.g., 10 words) and definitions belong to SF, denoted as ListofSLWords
Read IAA minimum score (threshold)
LingoGap: is a developed crowdsourcing tool (webiste)
Identify a list of crowd workers (w_1, w_2, ..., w_n) as L_{cw}
Identify a linguistic expert as Exp
The requester performs the following:
1) constructing the input SL and TL datasets in a specific SF
2) creating the crowdsourcing tasks.
3) monitoring tasks to control contribution quality.
4) validating and exporting crowdsourced data.
By LingoGap, the requester creates a crowdsourcing task (denoted as T) using the source words in ListofSLWords
Alpha: Krippendorff's Alpha denotes to the inter-annotator agreement (IAA)
The list of workers in (L_{cw}) conduct the task (T) and save their answers as (A_1)
yes
Alpha(A_1) >= threshold?
no
Register L_{cw} workers as HighQltyWorkers
i= n -1
Combination fun. is the mathematical combination function (Brualdi 2004)
Select a set (denoted as S_i) of all possible lists of workers from L_{cw} using the combination function, each is a list of size (i)
k= 0
Identify a list of workers at index k in S_i as (L_k)
Memeber s selection from L_{cw}
L_k and Exp conduct the task (T) and save their answers as (A_{ik})
k ++
no
Alpha(A_{ik}) >= threshold?
yes
i - -
no
k == S_i [length]?
yes
no
i == 0?
yes
Identifying the next combination of workers
Register the workers of L_{cw} as owQltyWorkers
end
Register (L_{cw} - L_{ik}) workers as LowQltyWorkers
Selecting low quality workers

FIGURE 1
Crowd filtering using alpha.

4.3 Step 3: task validation

We validate the crowdsourced gaps and words in two subsequent phases. First, *native-speaking crowd workers* perform data validation. We employ Alpha to measure IAA and filter out responses with low agreement. Second, a *linguistic expert* reviews the responses with low IAA identified in the first phase.

4.3.1 Crowd-based validation

A group of proficient, native-speaking crowd workers—those involved in the contribution collection described in Section 4.2.2—participate in a mutual validation process. In this process, each group cross-validates the contributions of another group, with the IAA scores across participants serving as the basis for evaluation.

TABLE 2 Guidelines for the experiment of English to Arabic (described in Section 5.2).

What will you be asked to do?	You will be asked to inspect 35 English food words and evaluate them by selecting one of two alternatives for each word: whether Arabic has an equivalent meaning or it is a lexical gap.
What is a lexical gap?	A lexical gap is a word with a distinct meaning that is missing from the vocabulary of a language.
What is an example of a lexical gap?	The English term "ham sandwich," referring to a sandwich filled with sliced ham, is a lexical gap in Arabic due to cultural differences. In many Arabic-speaking cultures, ham is not commonly consumed because of cultural and religious considerations, and sandwiches containing ham are not typically found in Arabic cuisine. Similarly, the Arabic word سحور meaning "the meal that Muslims eat before dawn during the month of Ramadan," represents a gap in the English-speaking community.
What are the needed qualifications?	Our experiment is not restricted by the user's background, culture, skills, etc. The only requirement is that participants must be native Arabic speakers with a good command of the English language. We prefer participants to have linguistic knowledge.
Are there any restrictions?	The only restriction is that you are not allowed to use machine translation, such as Google Translate, for translating word definitions. All possible meanings in Arabic are presented in a list in the answer section.
How long will the experiment take?	The maximum duration of the experiment is about 60 min. Please try to be as accurate as possible.
Tips:	We encourage you to (1) use a dictionary if you are unsure about your answer, and (2) search for an image of the English word on Wikipedia if you do not understand the English definition.
Will your data be processed anonymously?	The data collected will be kept strictly confidential; all responses will be stored and processed anonymously.
How will you indicate your consent?	By clicking "I consent, begin the study" below, you acknowledge that you speak English and Arabic, that you are at least 18 years old, and that you give your consent to participate in this study. If you do not intend to give consent, click "I do not consent; I want to withdraw from this study." Even after giving consent, you have the right to stop and withdraw from the experiment at any time without providing a reason. You can withdraw from the study at any time simply by closing the browser.
Contact person	If you have any questions, please contact the task requester (Hadi Khalilia) via email at hadi.khalilia@unitn.it.

IAA is widely used to assess the reliability of crowdsourced annotations in computational linguistics (Artstein and Poesio, 2008). Statistical measures such as Cohen's Kappa (Warrens, 2011) and Krippendorff's Alpha (Krippendorff, 2011) are commonly employed to evaluate consistency among annotators. Alpha is particularly versatile, accommodating various data types—nominal, ordinal, interval, and ratio—whereas Cohen's Kappa is most appropriate for nominal data and pairwise agreement (Powers, 2012).

In our crowdsourcing framework, which involves two or more annotators per item, we adopt Alpha to measure IAA. This approach enables us to systematically identify and exclude participants whose annotations lead to low agreement, ensuring that only items lacking sufficient consensus (with less than 100% IAA) are escalated for expert review.

To validate the data collected from an initial group of crowd workers (G_1: w_1, w_2, w_3), a second group of native-speaking crowd workers (G_2: w'_1, w'_2, w'_3) is engaged via the LingoGap platform. A visual overview of this procedure is provided in the flowchart in Figure 2. The validation process begins with reading the source and target language items, their definitions, and the responses collected from G_1. Inter-annotator agreement is then measured using Alpha. If the Alpha exceeds a predefined threshold, the G_1 workers are deemed high-quality, and only the items with disagreement are forwarded to a linguistic expert (*Exp*) for further evaluation.

If Alpha falls below the threshold, indicating low agreement, a new crowdsourcing task is launched through LingoGap. Subsets of workers from both G_1 and G_2 are formed using various combinations (Brualdi, 2004), ranging from individual workers to the full group. These subsets repeat the annotation task, after which Alpha is recalculated. If a subset's Alpha exceeds the threshold, the corresponding G_1 workers are classified as high-quality; those not included in such subsets are labeled as low-quality. An illustrative example is provided below.

Example

Consider two groups of crowd workers: $G_1 = \{w_1, w_2, w_3\}$ and $G_2 = \{w'_1, w'_2, w'_3\}$, along with a linguistic expert (*Exp*).

First, the IAA is assessed by computing the Alpha value for G_1. If Alpha meets the required threshold, all G_1 workers are deemed high-quality, and any of their responses with less than 100% agreement are sent to *Exp* for validation.

If Alpha falls below the threshold, the process continues by forming combinations of two workers from G_1 with one from G_2, producing 9 groups (e.g., $\{w_1, w_2, w'_1\}, \{w_1, w_3, w'_1\}, \ldots$). If any combination meets the threshold, the excluded G_1 worker is marked low-quality, and their partial-agreement responses are sent to *Exp*.

If none pass, the process uses one worker from G_1 with two from G_2, forming another 9 combinations (e.g., $\{w_1, w'_1, w'_2\}, \{w_2, w'_1, w'_3\}, \ldots$). If any such combination exceeds the threshold, the remaining G_1 workers are marked low-quality, and their responses with less than 100% IAA are sent to *Exp*.

If all these also fail, the task is repeated using only G_2. If their Alpha meets the threshold, all G_1 workers are considered low-quality. Otherwise, both groups are marked low-quality, and all G_1 responses are sent to *Exp*, concluding the validation.

4.3.2 Expert-based verification

A bilingual expert reviews responses from the Crowd-Based Validation stage that did not achieve 100% IAA. A spreadsheet is prepared containing the following columns: *Worker-ID*, *Source Lemma*, *Source Gloss*, and *Worker's Answer* (categorized as "*Lexical Gap*," "*Equivalent Word*," or "*Do not know*").

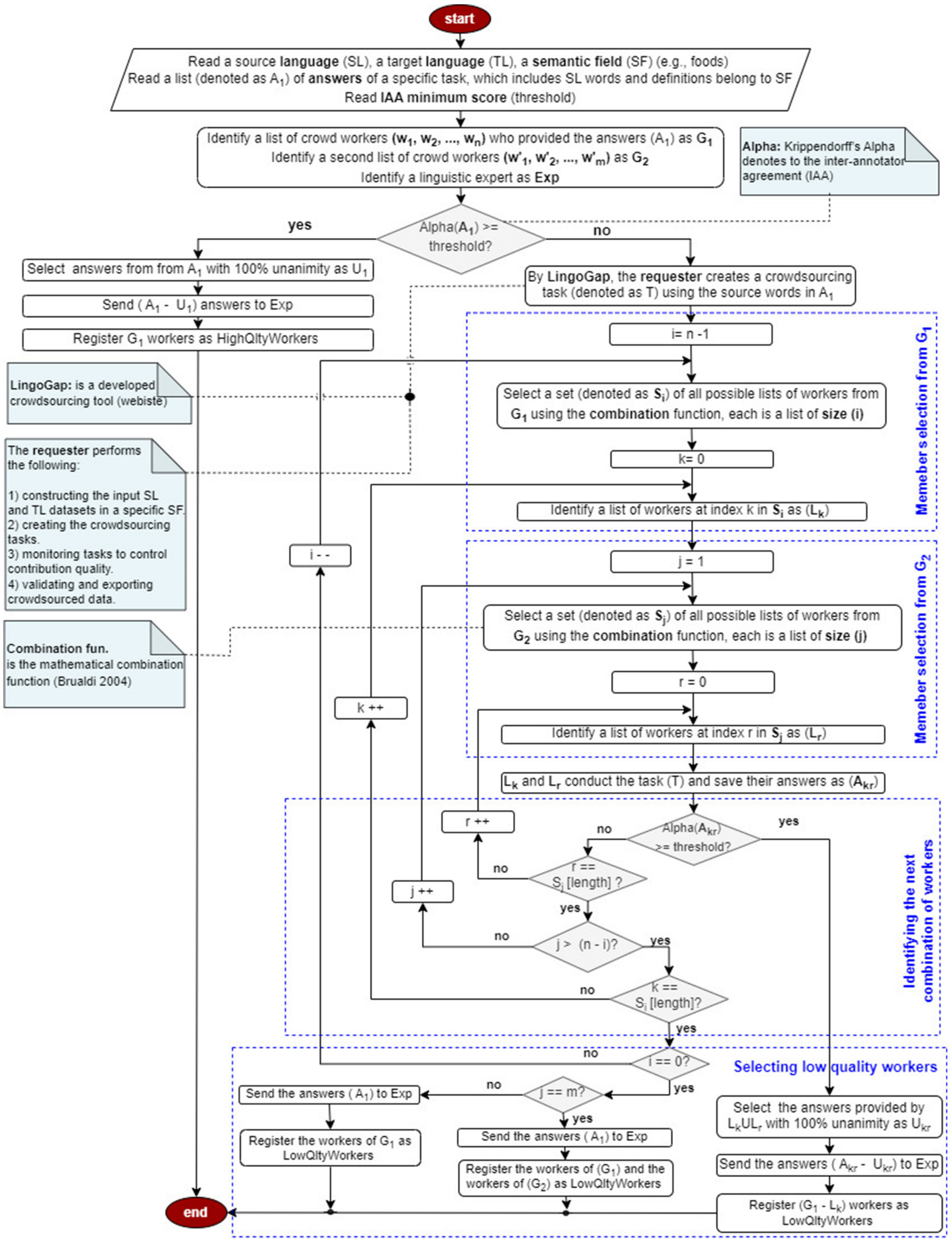

start
Read a source language (SL), a target language (TL), a semantic field (SF) (e.g., foods)
Read a list (denoted as A1) of answers of a specific task, which includes SL words and definitions belong to SF
Read IAA minimum score (threshold)
Identify a list of crowd workers (w1, w2, ..., wn) who provided the answers (A1) as G1
Identify a second list of crowd workers (w'1, w'2, ..., w'm) as G2
Identify a linguistic expert as Exp
Alpha: Krippendorff's Alpha denotes to the inter-annotator agreement (IAA)
Alpha(A1) >= threshold?
yes
no
Select answers from from A1 with 100% unanimity as U1
Send (A1 - U1) answers to Exp
Register G1 workers as HighQltyWorkers
By LingoGap, the requester creates a crowdsourcing task (denoted as T) using the source words in A1
LingoGap: is a developed crowdsourcing tool (webiste)
The requester performs the following:
1) constructing the input SL and TL datasets in a specific SF.
2) creating the crowdsourcing tasks.
3) monitoring tasks to control contribution quality.
4) validating and exporting crowdsourced data.
Combination fun. is the mathematical combination function (Brualdi 2004)
i= n -1
Select a set (denoted as Si) of all possible lists of workers from G1 using the combination function, each is a list of size (i)
k= 0
Identify a list of workers at index k in Si as (Lk)
Memeber s selection from G1
i - -
k ++
j = 1
Select a set (denoted as Sj) of all possible lists of workers from G2 using the combination function, each is a list of size (j)
r = 0
Identify a list of workers at index r in Sj as (Lr)
Memeber selection from G2
Lk and Lr conduct the task (T) and save their answers as (Akr)
Alpha(Akr) >= threshold?
yes
no
r ++
r == Sj [length] ?
no
yes
j ++
j > (n - i)?
no
yes
k == Si [length]?
no
yes
Identifying the next combination of workers
i == 0?
no
yes
j == m?
no
yes
Selecting low quality workers
Send the answers (A1) to Exp
Register the workers of G1 as LowQltyWorkers
Send the answers (A1) to Exp
Register the workers of (G1) and the workers of (G2) as LowQltyWorkers
Select the answers provided by LkULr with 100% unanimity as Ukr
Send the answers (Akr - Ukr) to Exp
Register (G1 - Lk) workers as LowQltyWorkers
end

FIGURE 2
Crowdsourced data validation using alpha.

Additionally, 10% of the entries in the spreadsheet—randomly selected from those with 100% IAA—are included as a “sanity check” to assess the expert’s consistency. The expert is then tasked with the following:

- *Equivalent meanings*: Evaluate the target language words provided by crowd workers and mark them as correct or incorrect. If a word is deemed incorrect, the expert supplies the correct equivalent or identifies it as a lexical gap.
- *Lexical gaps*: Confirm or reject the crowd’s classification of source words as lexical gaps. If a word is incorrectly marked as a gap, the expert provides the appropriate equivalent.
- *Do not know*: For source words marked as “*Do not know*,” the expert determines whether they represent lexical gaps or provides suitable equivalents in the target language.

5 Implementation and evaluation

In this section, we present the practical implementation and evaluation of our crowdsourcing methodology via the LingoGap platform. We begin by outlining the platform’s technical and functional features. Subsequently, we describe two case studies within the food semantic domain, involving diverse language pairs—English–Arabic and Indonesian–Banjarese—selected for their cultural and lexical variation. Finally, we evaluate the quality of the crowdsourced data by comparing it with annotations produced by large language models (LLMs), offering insights into the respective strengths and limitations of human and machine contributions in identifying lexical gaps.

5.1 LingoGap platform

To operationalize our crowdsourcing methodology, we developed LingoGap, a custom-built, web-based platform for collecting, managing, and validating lexical diversity data through structured micro-tasks. Developed using Java (JSP), JavaScript, and MySQL, LingoGap supports both administrative (requester) and worker roles, and it can function as a standalone system or integrate with commercial crowdsourcing platforms such as Prolific. The platform features two primary interfaces: the *requester interface* for researchers and the *worker interface* for crowd contributors.

Requester Interface[10]: The interface includes multiple tabs for task management. Using the “*Experiments*” tab, the requester can import SF datasets created in Section 4.1, configure tasks by specifying the SL and TL, and select a subset of words (e.g., 35 terms) from the full SL dataset displayed in a data table. The requester can also monitor task execution by defining ACQs and recording the completion times of crowd workers’ responses. Through the “*Source Words*” and “*Target Words*” tabs, lexical entries for both languages can be loaded and managed. Additionally, the requester can upload customized task instructions to enhance clarity. Requesters may import guideline spreadsheets into the LingoGap database. These spreadsheets contain configurable prompts (e.g., explanations of lexical gaps, task restrictions), expected response formats, and ethical considerations (e.g., anonymity, withdrawal rights), enabling the creation of flexible instruction templates tailored to the specific source and target languages. For instance, the *guidelines* used in our English-Arabic case study (Section 5.2) are detailed in Table 2.

Worker interface: Workers are presented with SL words and complete a multi-step evaluation for each word: (1) Gap Assessment—Determine whether the meaning exists in the TL. If not, the item is marked as a lexical gap; (2) Match Selection—If the meaning exists, select the equivalent from a provided list of TL words. For instance, Figure 3 shows a list of Arabic words for the selection of a food-related term in the example of exploring “*banana*” in Section 5.2; and (3) Custom Entry—If no suitable match is found, manually input a TL term along with a gloss. This multi-step format, *conceptually* inspired by the structured reasoning principle of Chain-of-Thought (CoT) approaches (Lin et al., 2024), encourages annotators to follow a clear, stepwise decision process that enhances task comprehension.

5.2 Case study on food terminology across English and Arabic

This case study investigates lexical diversity in the food domain across English and Arabic using the crowdsourcing methodology described in Section 4. The study was conducted in two phases: English-to-Arabic and Arabic-to-English, following the same structured workflow.

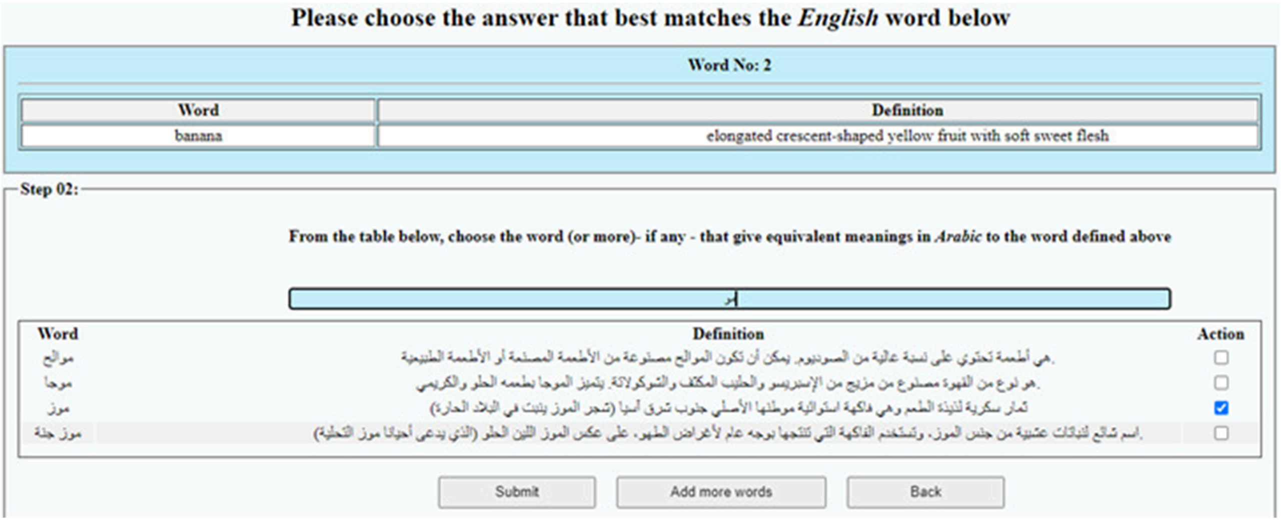
5.2.1 Study setup

5.2.1.1 Task generation

Two lexical datasets were developed: one in English (2,364 terms) and the other in Arabic (1,607 terms). The English food terms were extracted from the UKC using the Exporter tool, a built-in UKC management service designed to extract data from the UKC. The tool was used to collect English words belonging to a specific semantic field. It takes one or more concepts (e.g., food, nutrient) that represent a given semantic field (e.g., food) and the UKC resource as inputs, and then retrieves all lexicalizations associated with those concepts from the corresponding language lexicon in the UKC.

For Arabic, due to the lack of comprehensive lexical databases—particularly those containing food-related terms (e.g., UKC, WordNet)—a custom semantic filtering approach was employed. This method, based on AraBERT embeddings, extracted food-relevant terms from digital dictionaries (e.g., Almaany Dictionary) using cosine similarity. As illustrated in Figure 4, the method involves three main steps. In *Step 1*, a digital dictionary and food-related definitions (including example terms) are provided as inputs, which are then preprocessed by segmenting paragraphs into sentences. In *Step 2*, the sentences are transformed into vector representations using AraBERT. In *Step 3*, dictionary vectors similar to the centroid food vector are clustered based on cosine similarity with an 0.85 threshold, then converted back into text sentences and compiled in a spreadsheet. For clarity and ease of understanding, *English*

[10] http://lingogap.disi.unitn.it/admin.jsp



Please choose the answer that best matches the *English* word below

Word No: 2

Word	Definition
banana	elongated crescent-shaped yellow fruit with soft sweet flesh

Step 02:

From the table below, choose the word (or more)- if any - that give equivalent meanings in *Arabic* to the word defined above

مو

Word	Definition	Action
موالح	هي أطعمة تحتوي على نسبة عالية من الصوديوم. يمكن أن تكون الموالح مصنوعة من الأطعمة المصنعة أو الأطعمة الطبيعية.	☐
موجا	هو نوع من القهوة مصنوع من مزيج من الإسبريسو والحليب المكثف والشوكولاتة. يتميز الموجا بطعمه الحلو والكريمي.	☐
موز	ثمار سكرية لذيذة الطعم وهي فاكهة استوائية موطنها الأصلي جنوب شرق آسيا (شجر الموز ينبت في البلاد الحارة)	☑
موز جنة	اسم شائع لنباتات عشبية من جنس الموز، وتستخدم الفاكهة التي تنتجها بوجه عام لأغراض الطهو. على عكس الموز اللين الحلو (الذي يدعى أحيانا موز التحلية).	☐

FIGURE 3
Worker's GUI showing the step of match selection using LingoGap.

examples were used for the input and output demonstrations instead of Arabic ones. A custom Python script implementing this AraBERT-based methodology was developed to collect Arabic food-related words from the dictionary. The script is available on GitHub.[11]

5.2.1.2 Crowd selection

Native speakers proficient in the target language were recruited as volunteers: 12 Arabic native speakers with English fluency and 12 English native speakers with Arabic proficiency. All participants were university-educated. Each group was divided into four teams (G1–G4) of three members. A proficiency test and Alpha–based filtering methodology (depicted in Figure 1) were applied to identify high-quality contributors. For instance, 12 of 14 students passed the test successfully in the experiment of (English → Arabic), and for the application the filtering methodology in the same experiment, four tasks-each comprising 10 questions-were conducted . G_1 completed the first task, G_2 the second, G_3 the third, and G_4 the fourth, with all tasks performed by a linguistic expert. As a result, one student from G_1 was replaced, while G_2, G_3, and G_4 remained unchanged.

5.2.1.3 Contribution collection and quality control

Using the LingoGap platform and the guidelines outlined in Table 2, each group completed micro-tasks to evaluate source language terms and determine whether equivalent terms existed in the target language or whether a lexical gap was present. Three response options were available: equivalent term, lexical gap, or "do not know." Tasks included automated logging of completion times, ACQs to ensure attentiveness, and the exclusion of responses with anomalously short or long completion times. For example, in the *English* → *Arabic* experiment, for each English word, LingoGap displayed the term and asked crowd workers whether an equivalent existed in Arabic. If confirmed, LingoGap presented a list of Arabic food-related terms for selection, as illustrated with the term *banana* in Figure 3. If no equivalent was confirmed, the response was recorded as a lexical gap. If the equivalent term was not listed, the crowd was allowed to enter it manually via text input.

In the *English* → *Arabic* setting, each group processed subsets of the 2,364 English words against the full Arabic dataset. Groups G_1 and G_2 each worked on 245 English words distributed across seven crowdsourcing tasks covering *alcoholic drinks, pizza, salads, dairy products, rice, bread*, and *fruits*. Groups G_3 and G_4 worked on 945 and 929 English words, respectively, focusing on *soups, vegetables, cakes, meats, sandwiches*, and *desserts*.

In the *Arabic* → *English* direction, overlapping terms identified as equivalents in the (English → Arabic) experiment were excluded, resulting in 906 unique Arabic terms. The groups worked with these filtered Arabic terms evaluated against the full English dataset. Group G_1 was assigned 245 Arabic words and G_2 241 words, with both groups covering tasks on *sweets, rice meals, soups, vegetables, meats*, and *sandwiches*. Groups G_3 and G_4 each addressed 210 Arabic words across six tasks focusing on *drinks, pizza, salads, dairy products, bread*, and *fruits*.

Pilot tasks were conducted to refine the task design. Four pilot tasks were used to calibrate task parameters including the number of questions, number of ACQs, and task duration. The final configuration was set to 35 questions, 3 ACQs, and 60 minutes per task. Based on pilot task outcomes, three ACQs were adopted in all subsequent tasks. For instance, in Task 6 of the *English* → *Arabic* experiment conducted by G_2, responses from worker w'_1 were excluded due to a failed ACQ. Responses from w'_2 and w'_3, with an Alpha of 0.922, were retained (see Table 3). Additionally, outlier responses were filtered based on completion time per word. Among 68 tasks, 19 responses were discarded due to unusually fast completion times (3–6 seconds), compared to the average range of 80–120 seconds per task.

5.2.1.4 Task validation

For both the *English* → *Arabic* and *Arabic* → *English* experiments, we employed the methodology illustrated in Figure 2 to validate

[11] https://github.com/HadiPTUK/developed_scripts/blob/main/arabic_food_terms_script.py

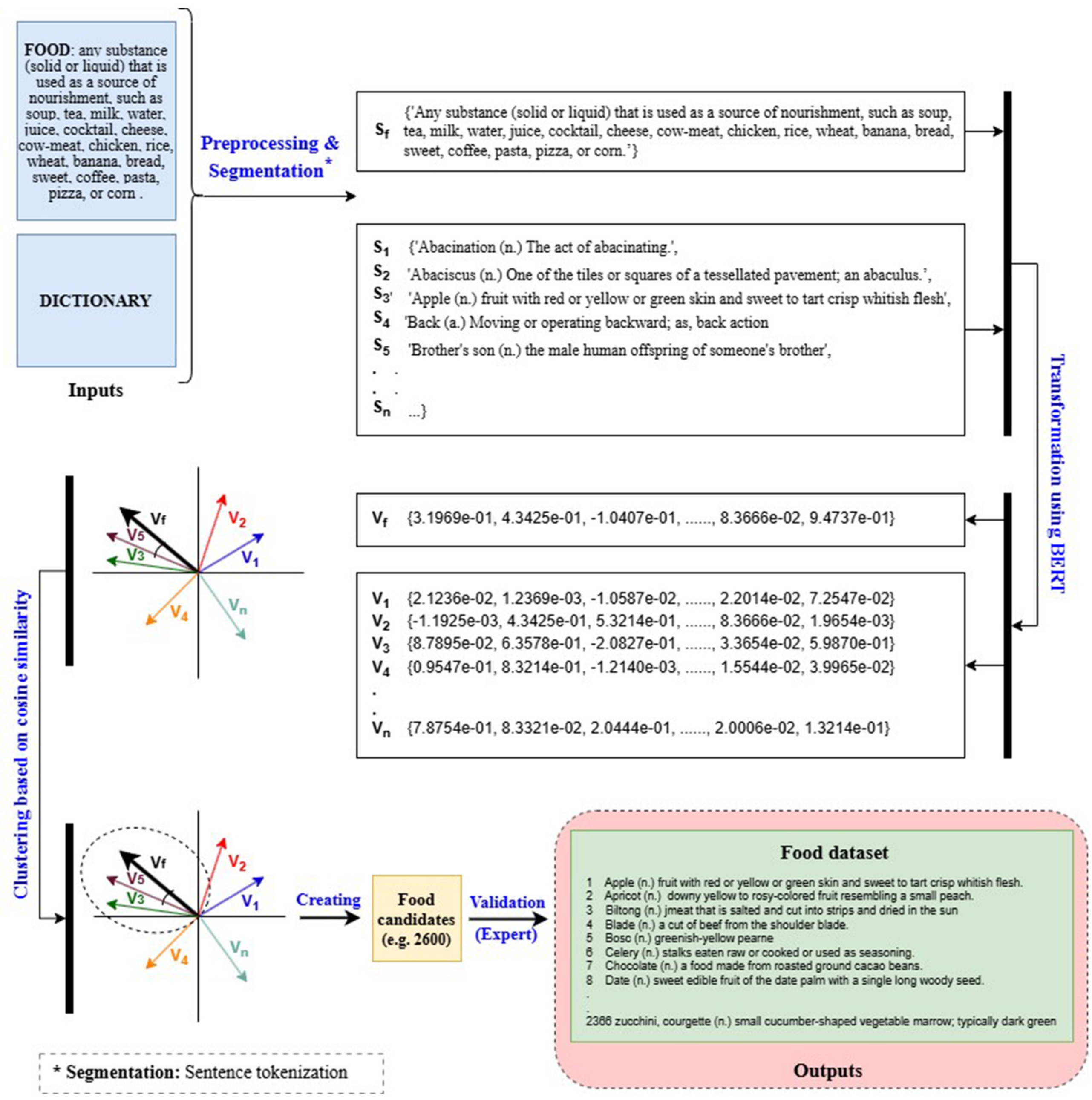

FOOD: any substance (solid or liquid) that is used as a source of nourishment, such as soup, tea, milk, water, juice, cocktail, cheese, cow-meat, chicken, rice, wheat, banana, bread, sweet, coffee, pasta, pizza, or corn .
DICTIONARY
Inputs
Preprocessing & Segmentation*
Sf {'Any substance (solid or liquid) that is used as a source of nourishment, such as soup, tea, milk, water, juice, cocktail, cheese, cow-meat, chicken, rice, wheat, banana, bread, sweet, coffee, pasta, pizza, or corn.'}
S1 {'Abacination (n.) The act of abacinating.',
S2 'Abaciscus (n.) One of the tiles or squares of a tessellated pavement; an abaculus.',
S3' 'Apple (n.) fruit with red or yellow or green skin and sweet to tart crisp whitish flesh',
S4 'Back (a.) Moving or operating backward; as, back action
S5 'Brother's son (n.) the male human offspring of someone's brother',
Sn ...}
Transformation using BERT
Vf {3.1969e-01, 4.3425e-01, -1.0407e-01, ..., 8.3666e-02, 9.4737e-01}
V1 {2.1236e-02, 1.2369e-03, -1.0587e-02, ..., 2.2014e-02, 7.2547e-02}
V2 {-1.1925e-03, 4.3425e-01, 5.3214e-01, ..., 8.3666e-02, 1.9654e-03}
V3 {8.7895e-02, 6.3578e-01, -2.0827e-01, ..., 3.3654e-02, 5.9870e-01}
V4 {0.9547e-01, 8.3214e-01, -1.2140e-03, ..., 1.5544e-02, 3.9965e-02}
Vn {7.8754e-01, 8.3321e-02, 2.0444e-01, ..., 2.0006e-02, 1.3214e-01}
Clustering based on cosine similarity
Vf
V5
V3
V2
V1
V4
Vn
Creating
Food candidates (e.g. 2600)
Validation (Expert)
Food dataset
1 Apple (n.) fruit with red or yellow or green skin and sweet to tart crisp whitish flesh.
2 Apricot (n.) downy yellow to rosy-colored fruit resembling a small peach.
3 Biltong (n.) jmeat that is salted and cut into strips and dried in the sun
4 Blade (n.) a cut of beef from the shoulder blade.
5 Bosc (n.) greenish-yellow pearne
6 Celery (n.) stalks eaten raw or cooked or used as seasoning.
7 Chocolate (n.) a food made from roasted ground cacao beans.
8 Date (n.) sweet edible fruit of the date palm with a single long woody seed.
2366 zucchini, courgette (n.) small cucumber-shaped vegetable marrow; typically dark green
Outputs
* Segmentation: Sentence tokenization

FIGURE 4
Flowchart of the semantic-field filtering method to collect food words from a digital dictionary.

crowdsourced data from groups G_1, G_2, G_3, and G_4. IAA, measured using Alpha, generally exceeded the acceptance threshold of 0.70. However, in the *English* → *Arabic* setting, two exceptions were noted: Task 5 (by G_2) and Task 17 (by G_3) initially scored Alpha values of 0.59 and 0.62, respectively. These tasks were subsequently reassigned—Task 5 to workers from G_1 and G_2, and Task 17 to workers from G_3 and G_4—resulting in improved Alpha scores of 0.89 and 0.82. Workers w_3' (from G_2) and w_1' (from G_3) were replaced in later tasks due to inaccuracies that contributed to disagreement.

In contrast, all tasks in the *Arabic* → *English* experiment achieved Alpha values above 0.70. Despite satisfactory agreement levels, all newly introduced words and responses with less than 100% IAA were submitted for expert validation in both experiments. In the *English* → *Arabic* direction, this included 88 English words. The expert confirmed the validity of all proposed Arabic words and suggested alternatives for two lexical gaps. For example, خبز شراك "*khubz shrak*" was recommended for "*chapatti*" instead of leaving a gap, as initially proposed by G_3 in Task 12.

TABLE 3 Crowdsourced data summary by G_1 and G_2 in Arabic.

Task	Gaps		Words		New concepts		Alpha	
	G1	G2	G1	G2	G1	G2	G1	G2
1	19	32	16	3	1	2	0.76	0.72
2	34	30	1	5	0	1	1.00	0.73
3	34	26	1	9	3	1	1.00	0.95
4	19	29	16	6	2	1	0.80	0.72
5	27	21	8	14	2	0	0.71	0.89
6	26	21	9	14	2	3	0.85	0.92
7	21	26	14	9	3	3	0.92	0.77
Total	180	185	65	60	13	11		
Avg.							0.86	0.81

Similarly, in the *Arabic* → *English* experiment, the expert reviewed 79 Arabic words with non-unanimous responses. All new English translations were validated, though four Arabic words were identified as lexical gaps. For instance, مهلبية—"*a dessert made from milk, starch, and sugar*"—was classified as a lexical gap, with the previously suggested equivalent "*pudding*" (provided by G_2 in Task 4) deemed too broad in semantic scope.

5.2.2 Study results

This section presents the results of two crowdsourcing experiments-English-to-Arabic and Arabic-to-English mappings-conducted to explore lexical diversity between English and Arabic in the domain of food-related terms.

Across the *English* → *Arabic* experiment, participants identified 1,532 lexical gaps, 832 equivalent words, and 100 new Arabic words that were not present in the original Arabic input dataset. The annotation process achieved a high inter-annotator agreement, with an average Alpha score of 0.84. Detailed information for the Arabic crowdsourced data are presented in Tables 3, 4.

Conversely, in the *Arabic* → *English* direction, the experiment yielded 608 lexical gaps, 298 equivalent words, and 49 new English words not found in the original English input. This task also demonstrated strong annotator consistency, with an average Alpha score of 0.85. Additional information on the English crowdsourced data is provided in Tables 5, 6.

The resulting datasets are publicly available via the DataScientia repository. The *English* → *Arabic* dataset can be accessed at[12], and the *Arabic* → *English* dataset is available at[13].

5.2.3 Lexical diversity evaluation: overlap-based metric

Several shared meaning overlaps have been found between language pairs. For a given domain d and two languages l_A and l_B, the formula below calculates the similarity of the two languages in terms of the overlap of lexicalised concepts from that domain, where $\text{LexCons}(d, l)$ stands for the set of domain concepts that are lexicalized by the language l.

$$\text{overlap}(d, l_A, l_B) = \frac{|\text{LexCons}(d, l_A) \cap \text{LexCons}(d, l_B)|}{\max(|\text{LexCons}(d, l_A)|, |\text{LexCons}(d, l_B)|)} \quad (1)$$

Figure 5 shows the overlaps between English and Arabic over the food domain. For example, the intersection of English and Arabic languages gives a shared coverage of 46.8%. The number of lexicalisations in English is 2,413 (2,364 words in the English input dataset and 49 new words, which are missing from the English input dataset), and in Arabic is 1,707 (1,607 words in the Arabic input dataset and 100 new words). Also, 1,130 of these lexical units (832 words were explored in the experiment of *English* → *Arabic*, 298 words were identified in the experiment of *Arabic* → *English*) are included in both languages. For example, Equation 1 calculates the overlap between English and Arabic (both represented in ISO 639-3 as "eng" and "arb," respectively) in the food domain (F) as follows:

$$\begin{aligned}&\text{overlap}(F, \text{eng}, \text{arb})\\ &= \frac{|\text{LexCons}(F,\text{eng}) \cap \text{LexCons}(F,\text{arb})|}{\max(|\text{LexCons}(F,\text{eng})|,|\text{LexCons}(F,\text{arb})|)}\end{aligned} \quad (2)$$

$$\text{overlap}(F, \text{eng}, \text{arb}) = \frac{1130}{\max(2413, 1707)} = \frac{1130}{2413} = 46.8\% \quad (3)$$

We find this overlap is lower than our initial expectations on language variations. Language experts justify such differences with two major factors: linguistic and religious influence (Albala, 2011; Armanios and Ergene, 2018). By linguistic influence, we refer to the etymological origin and borrowing of the language, which

[12] https://ds.datascientia.eu/community/public/projects/aef86b5b-5742-4529-9c46-71b5142f91bd

[13] https://ds.datascientia.eu/community/public/projects/db9ec453-4775-43d6-a89a-85a9600b9781

TABLE 4 Crowdsourced data summary by G_3 and G_4 in Arabic.

Task	Gaps		Words		New concepts		Alpha	
	G3	G4	G3	G4	G3	G4	G3	G4
1	29	23	6	12	1	1	0.81	0.88
2	21	33	14	2	2	2	0.73	0.79
3	24	20	11	15	2	1	0.79	0.92
4	17	21	18	14	1	2	0.92	0.91
5	19	15	16	20	2	2	0.81	0.80
6	24	27	11	8	0	0	0.83	0.95
7	18	22	17	13	0	0	0.79	0.88
8	24	20	11	15	3	1	0.77	0.87
9	24	13	11	22	2	2	0.71	0.92
10	14	24	21	11	2	0	0.76	0.96
11	20	21	15	14	0	2	0.87	0.89
12	25	16	10	19	2	2	0.78	0.88
13	29	23	6	12	1	2	0.93	0.84
14	30	26	5	9	0	1	0.86	0.91
15	23	27	12	8	2	2	0.92	0.90
16	32	28	3	7	0	1	0.78	0.79
17	18	16	17	19	1	3	0.62	0.89
18	24	22	11	13	2	1	0.91	0.88
19	14	19	21	16	1	2	0.80	0.92
20	22	21	13	14	2	3	0.87	0.73
21	19	22	16	13	1	0	0.89	0.88
22	24	23	11	12	1	2	0.91	0.72
23	17	20	18	15	0	3	0.81	0.81
24	19	21	16	14	1	3	0.85	0.85
25	24	24	11	11	3	2	0.87	0.87
26	23	22	12	13	0	0	0.90	0.84
27	15	6	20	13	3	1	0.83	0.86
Total	592	575	353	354	35	41		
Average							0.83	0.86

affects the lexicons. The two languages have distinct etymological origins. Arabic, a Semitic language, derives many food terms from ancient roots and influences from Persian, Turkish, and Indian cultures. English, a Germanic language, has borrowed food-related vocabulary from French, Italian, and other European languages over time. Secondly, the religion of the speaker community also affects the lexicon. In many Arabic-speaking regions, Islam significantly influences food practices. Halal dietary laws shape food culture, while pork and alcohol, common in some Western cuisines, are prohibited in the Arabic community. English-speaking countries have more religious diversity, which can allow for a broader range of food-related terms tied to various cuisines (Armanios and Ergene, 2018).

5.3 Case study on food terminology across Indonesian and Banjarese

To investigate lexical diversity in the food domain between Standard Indonesian[14] and Banjarese, we conducted two bidirectional experiments—matching from Indonesian to Banjarese and vice versa—using the crowdsourcing methodology described in Section 4.

[14] Throughout this paper, the term "Indonesian" refers specifically to "Standard Indonesian."

TABLE 5 Crowdsourced data summary by G_1 and G_2 in English.

Task	Gaps		Words		New concepts		Alpha	
	G1	G2	G1	G2	G1	G2	G1	G2
1	29	19	6	16	3	2	0.87	0.92
2	25	24	10	11	1	0	0.75	0.92
3	24	19	11	16	2	1	0.82	0.85
4	27	28	8	7	2	2	0.84	0.80
5	23	25	12	10	1	2	0.89	0.81
6	27	19	8	16	2	2	0.83	0.77
7	21	29	14	2	1	3	0.77	0.85
Total	176	163	69	78	12	12		
Avg.							0.82	0.85

TABLE 6 Crowdsourced data summary by G_3 and G_4 in English.

Task	Gaps		Words		New concepts		Alpha	
	G3	G4	G3	G4	G3	G4	G3	G4
1	21	26	14	9	0	2	0.81	0.91
2	18	26	17	9	2	3	0.89	0.95
3	32	22	3	13	3	1	0.72	0.88
4	15	24	20	11	3	2	0.88	0.87
5	21	22	14	13	4	2	0.87	0.85
6	15	27	20	8	2	1	0.92	0.81
Total	122	147	88	63	14	11		
Avg.							0.85	0.88

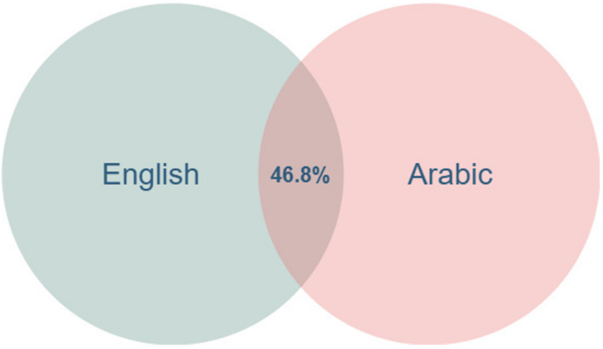


FIGURE 5
The overlap (percentage of shared lexicalizations) for English and Arabic languages.

5.3.1 Study setup

5.3.1.1 Task generation

We developed two lexical datasets comprising 1,448 Indonesian terms and 812 Banjarese terms, both obtained using semantic filtering methods. For the Indonesian dataset, we employed a similar approach to that used for Arabic (Figure 4). This method consists of three main steps. In *Step 1*, the Kamus Bahasa Indonesia (Tim Penyusun Kamus Pusat Bahasa, 2008) was used to extract all Indonesian terms. In *Step 2*, the terms obtained from *Step 1* were transformed into vector representations using IndoBERT (Koto et al., 2020). In *Step 3*, these vectors were compared to the vector representation of the term "food" using cosine similarity to identify terms most similar to the centroid. A custom Python script based on IndoBERT was developed to implement this methodology and collect food-related words in Indonesian from the Kamus Bahasa Indonesia dictionary. The script is available on GitHub.[15]

In the case of the Banjarese dataset, the absence of a high-quality dictionary or monolingual language model posed a significant challenge. To address this, we employed alternative resources—namely Word2Vec (Mikolov et al., 2013), Wiktionary, and Wikipedia—to compile the dataset through a three-step process. Figure 6 illustrates the methodological flowchart with examples, detailing the inputs, processing steps, and outputs. For clarity and ease of understanding, these examples are presented in *English*. In *Step 1*, a Banjarese corpus was created by extracting data from the NLLB dataset (Team et al., 2022) via NusaCrowd (Cahyawijaya et al., 2023). This corpus was then used to build and train a static

[15] https://github.com/HadiPTUK/developed_scripts/blob/main/indonesian_food_terms_script.py

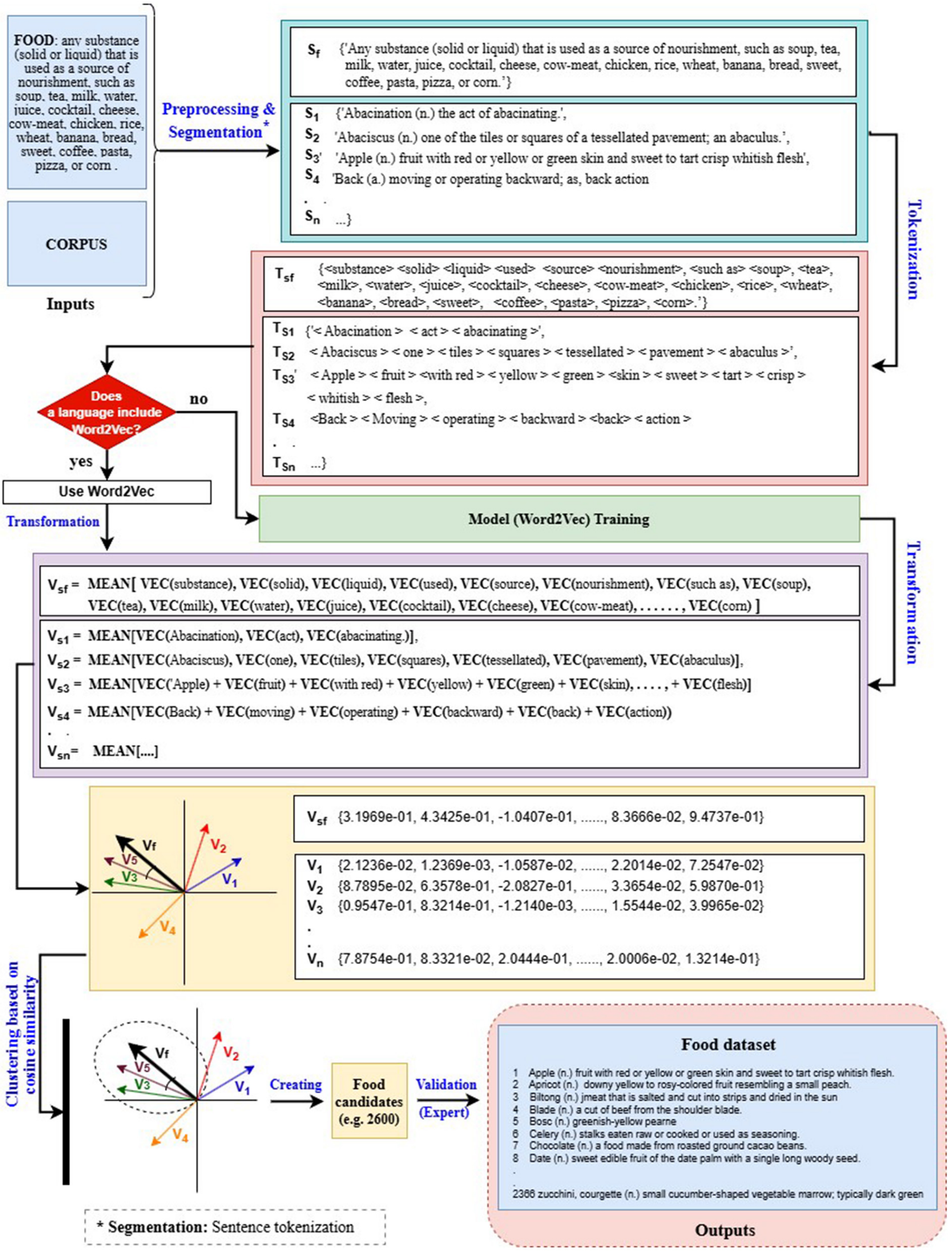

FOOD: any substance (solid or liquid) that is used as a source of nourishment, such as soup, tea, milk, water, juice, cocktail, cheese, cow-meat, chicken, rice, wheat, banana, bread, sweet, coffee, pasta, pizza, or corn .
CORPUS
Inputs
Preprocessing & Segmentation*
S_f {'Any substance (solid or liquid) that is used as a source of nourishment, such as soup, tea, milk, water, juice, cocktail, cheese, cow-meat, chicken, rice, wheat, banana, bread, sweet, coffee, pasta, pizza, or corn.'}
S_1 {'Abacination (n.) the act of abacinating.',
S_2 'Abaciscus (n.) one of the tiles or squares of a tessellated pavement; an abaculus.',
S_3 'Apple (n.) fruit with red or yellow or green skin and sweet to tart crisp whitish flesh',
S_4 'Back (a.) moving or operating backward; as, back action
S_n ...}
Tokenization
T_sf {<substance> <solid> <liquid> <used> <source> <nourishment>, <such as> <soup>, <tea>, <milk>, <water>, <juice>, <cocktail>, <cheese>, <cow-meat>, <chicken>, <rice>, <wheat>, <banana>, <bread>, <sweet>, <coffee>, <pasta>, <pizza>, <corn>.'}
T_S1 {'< Abacination > < act > < abacinating >',
T_S2 < Abaciscus > < one > < tiles > < squares > < tessellated > < pavement > < abaculus >',
T_S3 < Apple > < fruit > <with red > < yellow > < green > <skin > < sweet > < tart > < crisp > < whitish > < flesh >,
T_S4 <Back > < Moving > < operating > < backward > <back> < action >
T_Sn ...}
Does a language include Word2Vec?
no
yes
Use Word2Vec
Model (Word2Vec) Training
Transformation
V_sf = MEAN[VEC(substance), VEC(solid), VEC(liquid), VEC(used), VEC(source), VEC(nourishment), VEC(such as), VEC(soup), VEC(tea), VEC(milk), VEC(water), VEC(juice), VEC(cocktail), VEC(cheese), VEC(cow-meat), ... , VEC(corn)]
V_s1 = MEAN[VEC(Abacination), VEC(act), VEC(abacinating.)],
V_s2 = MEAN[VEC(Abaciscus), VEC(one), VEC(tiles), VEC(squares), VEC(tessellated), VEC(pavement), VEC(abaculus)],
V_s3 = MEAN[VEC('Apple) + VEC(fruit) + VEC(with red) + VEC(yellow) + VEC(green) + VEC(skin), ... , + VEC(flesh)]
V_s4 = MEAN[VEC(Back) + VEC(moving) + VEC(operating) + VEC(backward) + VEC(back) + VEC(action))
V_sn= MEAN[....]
Vf
V2
V5
V3
V1
V4
V_sf {3.1969e-01, 4.3425e-01, -1.0407e-01, ..., 8.3666e-02, 9.4737e-01}
V_1 {2.1236e-02, 1.2369e-03, -1.0587e-02, ..., 2.2014e-02, 7.2547e-02}
V_2 {8.7895e-02, 6.3578e-01, -2.0827e-01, ..., 3.3654e-02, 5.9870e-01}
V_3 {0.9547e-01, 8.3214e-01, -1.2140e-03, ..., 1.5544e-02, 3.9965e-02}
V_n {7.8754e-01, 8.3321e-02, 2.0444e-01, ..., 2.0006e-02, 1.3214e-01}
Clustering based on cosine similarity
Creating
Food candidates (e.g. 2600)
Validation (Expert)
Food dataset
1 Apple (n.) fruit with red or yellow or green skin and sweet to tart crisp whitish flesh.
2 Apricot (n.) downy yellow to rosy-colored fruit resembling a small peach.
3 Biltong (n.) jmeat that is salted and cut into strips and dried in the sun
4 Blade (n.) a cut of beef from the shoulder blade.
5 Bosc (n.) greenish-yellow pearne
6 Celery (n.) stalks eaten raw or cooked or used as seasoning.
7 Chocolate (n.) a food made from roasted ground cacao beans.
8 Date (n.) sweet edible fruit of the date palm with a single long woody seed.
2366 zucchini, courgette (n.) small cucumber-shaped vegetable marrow; typically dark green
Outputs
* Segmentation: Sentence tokenization

FIGURE 6
Flowchart of the semantic-field filtering method to collect food words from a corpus (e.g., Wikipedia).

word embedding model using Word2Vec. In *Step 2*, a custom Python script[16] was employed to extract a list of terms and their definitions from both Wiktionary and Wikipedia. These terms were subsequently transformed into vector representations using the Word2Vec model developed in *Step 1*. In *Step 3*, the vectorized terms were clustered around the term "food" and its definition, which served as the centroid. Cosine similarity was used to compute the distance between each term and the centroid, with a threshold set at 0.85. Terms exceeding this threshold were classified as food-related and added to the dataset.

5.3.1.2 Crowd selection

Native speakers of Banjarese and Indonesian with sufficient bilingual proficiency were recruited as volunteers and screened using a two-phase quality control process: a proficiency test followed by the methodology illustrated in Figure 1. The final pool included six Indonesian native speakers fluent in Banjarese and six Banjarese native speakers proficient in Indonesian. All participants were university students. Each group was divided into two teams (G1 and G2) of three members.

5.3.1.3 Contribution collection and quality control

As in the English-Arabic experiment, we used the LingoGap platform and followed the guidelines outlined in Table 2 to enable qualified crowd workers to match Indonesian and Banjarese terms through micro-tasks. In both directions (*Indonesian* → *Banjarese* and *Banjarese* → *Indonesian*), participants were assigned 45 terms and 4 ACQs per 90-minute task. This configuration was based on our pilot experiments for this study.

In the *Indonesian* → *Banjarese* experiment, 1,448 Indonesian terms were randomly distributed across 30 crowdsourcing tasks, as presented in Table 7, ensuring that no group focused exclusively on specific food categories. Each of Groups G1 and G2 completed 15 tasks. In the *Banjarese* → *Indonesian* experiment, 330 Banjarese terms—non-overlapping with their Indonesian equivalents—were similarly distributed across 8 tasks, as shown in Table 8.

5.3.1.4 Task validation

To ensure the reliability and quality of volunteer responses in both translation directions, we employed the validation methodology illustrated in Figure 2. A minimum agreement threshold of 0.7 was established and consistently exceeded by both participant groups. In the *Banjarese* → *Indonesian* direction, certain entries lacking full (100%) inter-annotator agreement were submitted to a linguistic expert for adjudication. For instance, the Banjarese word *rabuk*—meaning "*a type of ground meat*"—was identified as a lexical gap by some participants; the expert corrected this by providing the Indonesian equivalent *abon*. For further details, the corresponding datasets are available in the DataScientia repository: the Banjarese dataset[17] and the Indonesian dataset.[18]

5.3.2 Study results

This section presents the datasets generated from two crowdsourcing experiments involving food-related terms in Indonesian and Banjarese.

In the *Indonesian* → *Banjarese* experiment, crowd workers identified 750 lexical gaps, 507 equivalent terms, and 43 new Banjarese words not present in the original Banjarese input dataset. The annotation process demonstrated strong inter-annotator agreement, with an average Alpha score of 0.83. Additional details on the Banjarese crowdsourced data are provided in Table 7.

In the *Banjarese* → *Indonesian* direction, the experiment revealed 201 lexical gaps, 98 equivalent terms, and 30 new Indonesian words that were absent from the original Indonesian dataset. This task also exhibited good annotator consistency, with an average Alpha score of 0.83. Further information on the Indonesian crowdsourced data is available in Table 8.

5.3.3 Lexical diversity evaluation: overlap-based metric

This section examines the overlap between Indonesian and Banjarese in the food domain. As shown in Figure 7, the intersection of the Indonesian and Banjarese languages for a shared coverage of 40.9%. The number of lexicalisations in Indonesian is 1,478 (1,448 words in the original Indonesian input dataset and 30 new words), and in Banjarese is 855 (812 words in the original Banjarese input dataset and 43 new words). Also, 605 of these lexical units (507 words were explored in the experiment of *Indonesian* → *Banjarese*, 98 words were identified in the experiment of *Banjarese* → *Indonesian*) are included in both languages. For example, Equation 1 calculates the overlap between Indonesian and Banjarese (both represented in ISO 639-3 as "ind" and "bjn," respectively) in the food domain (F) as follows:

$$\text{overlap}(F, \text{ind}, \text{bjn}) \\ = \frac{|\text{LexCons}(F,\text{ind}) \cap \text{LexCons}(F,\text{bjn})|}{\max(|\text{LexCons}(F,\text{ind})|,|\text{LexCons}(F,\text{bjn})|)} \quad (4)$$

$$\text{overlap}(F, \text{ind}, \text{bjn}) = \frac{605}{\max(1478, 855)} = \frac{605}{1478} = 40.9\% \quad (5)$$

While both Indonesian and Banjarese are part of the Austronesian language family, they have different linguistic histories and influences. Banjarese has absorbed vocabulary from Dayak languages, Malay dialects, and other indigenous languages of Kalimantan, while Indonesian, as the national language, has been influenced by Malay, Javanese, Dutch, Arabic, and other foreign languages. Furthermore, these languages exist on different islands in Indonesia; Banjarese is located on the southern part of Borneo Island, and the Indonesian language 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. These historical and geographical influences leads to differences in vocabulary, including food terms.

[16] https://github.com/HadiPTUK/developed_scripts/blob/main/banjarese_food_terms_script.py

[17] https://ds.datascientia.eu/community/public/projects/c3d242b7-4ddc-419f-bffd-38d9f9760a05

[18] https://ds.datascientia.eu/community/public/projects/eacdb797-f8bf-45d4-8909-a4fd50ae8910

TABLE 7 Crowdsourced Banjarese data summary.

Task	Gaps		Words		New concepts		Alpha	
	G1	G2	G1	G2	G1	G2	G1	G2
1	30	36	12	6	2	1	0.86	0.83
2	29	35	13	6	1	3	0.81	0.83
3	23	35	18	7	2	1	0.82	0.80
4	19	41	23	1	1	1	0.82	0.84
5	22	42	19	2	2	0	0.83	0.79
6	20	35	21	9	3	0	0.91	0.81
7	22	32	20	9	2	2	0.83	0.82
8	25	34	17	8	1	1	0.82	0.85
9	11	26	32	14	0	3	0.82	0.82
10	21	28	21	13	1	2	0.82	0.81
11	19	32	25	10	0	1	0.85	0.83
12	11	32	31	10	1	1	0.85	0.83
13	12	28	30	13	2	3	0.82	0.87
14	7	30	34	11	2	3	0.88	0.84
15	8	5	34	38	1	0	0.82	0.82
Total	279	471	350	157	21	22		
Average							0.84	0.83

TABLE 8 Crowdsourced Indonesian data summary.

Task	Gaps		Words		New concepts		Alpha	
	G1	G2	G1	G2	G1	G2	G1	G2
1	26	25	15	17	4	3	0.82	0.87
2	28	29	14	12	3	4	0.85	0.81
3	28	28	12	13	5	4	0.82	0.84
4	30	7	11	4	4	3	0.81	0.81
Total	112	89	52	46	16	14		
Avg.							0.83	0.83

5.4 Quality of crowdsourced data: a comparison with LLMs annotation

To complement the crowdsourced annotations, we evaluated whether LLMs could perform the same lexical gap identification task, aiming to assess how effectively current LLMs capture cultural and semantic nuances typically recognized by native speakers. This section explores the use of LLMs as annotation tools in linguistic research, focusing on the construction of small, diversity-aware datasets that serve as benchmarks for evaluating the reliability of diversity-related data obtained through crowdsourcing. Specifically, we assess the translatability of lexical items to examine lexical diversity and semantic equivalence across source and target languages, using LLMs to capture nuanced cross-linguistic correspondences.

Recent studies underscore the growing effectiveness of LLMs, particularly GPT-based models, in text annotation tasks. For instance, GPT-3 and GPT-4 have been employed to annotate political tweets and identify propaganda (Ding et al., 2023; Hasanain et al., 2024), while Google's Gemini Pro has been used for structured data extraction in scientific texts (Sayeed et al., 2024). Comparative evaluations indicate that GPT models frequently outperform both crowdsourced annotators and domain experts in tasks such as political stance detection (Alizadeh et al., 2023; Törnberg, 2023). Building on this body of work, we utilize GPT-4, DeepSeek, and Gemini to annotate cross-linguistic data, with a focus on lexical diversity and semantic equivalence. GPT-4 is prioritized due to its demonstrated superior performance.

Our annotation process adapts the crowdsourcing methodology described in Section 4, incorporating two key modifications. *First*,

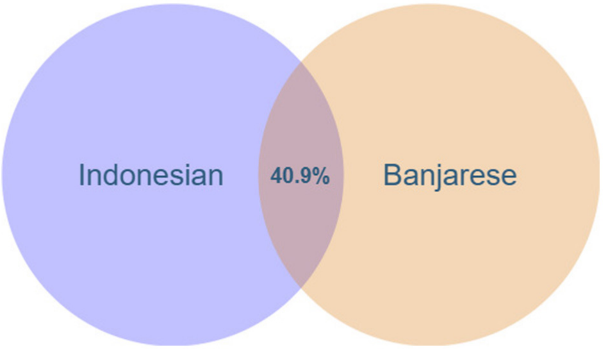


FIGURE 7
The overlap (percentage of shared lexicalizations) for Indonesian and Banjarese languages.

we replace crowdsourcing with an *Annotation Phase*, during which LLMs perform the annotation. Each task consists of (50)[19] SL lexical items within the specified SF, paired with TL candidates. The objective is to determine whether a TL equivalent exists. If it does, the TL entry is selected or proposed as a new term; if not, the SL item is marked as a lexical gap. Figure 8 presents the prompt template used. *Second*, we rely exclusively on expert validation to assess the accuracy of identified equivalents and lexical gaps. This method is applied in two food-domain case studies: *English–Arabic* and *Indonesian–Banjarese*. These language pairs were also examined using our crowdsourcing approach, as detailed in Sections 5.2, 5.3.

Each experiment was conducted using GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash via their web interfaces. The models were neither fine-tuned nor externally assisted, ensuring an authentic comparison with human annotators. Each query followed the same structure (as shown in Figure 8) and input format as the human task to maintain methodological consistency. Specifically, each LLM received the SL word and its definition (gloss) as input and was prompted to determine whether an equivalent existed in the TL.

As demonstrated in the two case studies, the results indicate that LLMs struggle to identify lexical gaps, particularly in low-resource and culturally diverse contexts. Three key findings emerged:

- *Hallucination of TL terms*: LLMs occasionally generated plausible but nonexistent TL words when no equivalent existed, suggesting overgeneralization.
- *Failure in culturally grounded semantics*: The models frequently misinterpreted culturally embedded terms (e.g., food items), generating translations that were overly generic or semantically inaccurate.
- *Lack of sociocultural awareness*: Unlike human annotators, LLMs were unable to infer when the absence of a TL equivalent reflected a deeper cultural or linguistic divergence.

While the LLM performed well on straightforward lexical mappings, it demonstrated limited sensitivity to subtle semantic distinctions and sociocultural contexts. These findings reinforce the paper's central argument that human-in-the-loop crowdsourcing remains essential for capturing linguistic diversity and validating lexical gaps. Although LLMs can assist with data generation and pre-filtering, they cannot yet replicate the nuanced reasoning of native speakers in cross-linguistic annotation tasks.

5.4.1 Case study: food terminology in English and Arabic

This case study investigates cross-linguistic lexical diversity between English and Arabic through two experiments. The first experiment (*English → Arabic*) examines how English food terms are lexicalized in Arabic. Fifty culturally specific English food terms were selected from a dataset of 2,364 terms compiled in Section 5.2. These terms were compared against an Arabic dataset of 1,607 words (also created in Section 5.2) using GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash. Each model performed zero-shot annotations, either identifying an Arabic equivalent or marking the term as a lexical gap.

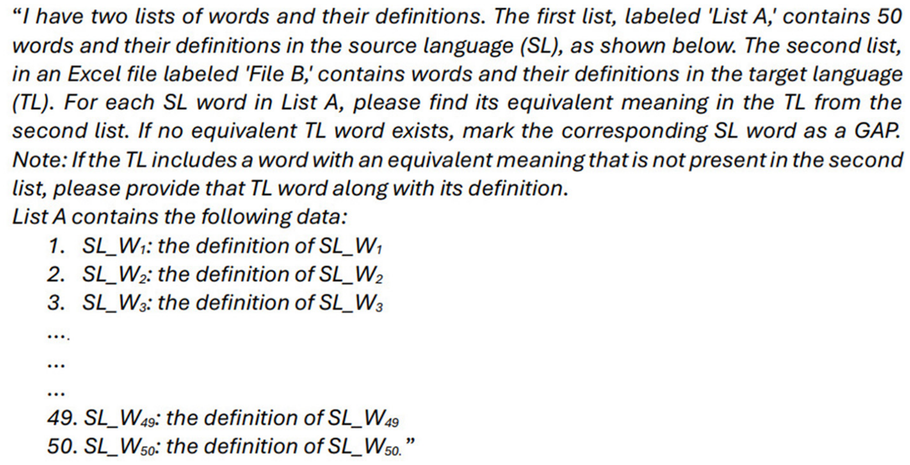
The second experiment (*Arabic → English*) reversed the direction of analysis. Fifty Arabic-specific food terms were selected and matched against the English dataset using the same language models. Each model annotated the Arabic terms by either identifying English equivalents or indicating lexical gaps.

Expert validation was conducted to ensure annotation accuracy. An assistant professor specializing in lexical semantics and a native speaker of Arabic validated the Arabic equivalents and lexical gaps. An Arabic-speaking university lecturer based in the United Kingdom reviewed the English annotations. Table 9 presents the models' annotation *accuracy*, defined as the number of correctly identified equivalents or gaps divided by the total number of source terms. Two common error types were observed: (1) providing *incorrect TL equivalents*, and (2) offering *literal translations* instead of correctly identifying lexical gaps.

For instance, in the *Arabic → English* experiment, the Arabic term الأبيضان "*water and yogurt*" was mistranslated by GPT-4o as "*the two whites*." Similarly, كبة "*a dish made of ground meat and rice or wheat*" was translated literally as *kibbeh* by all models, although the term lacks recognition in standard English lexicons.

In the *English → Arabic* experiment, *malt liquor*, meaning "*a strong lager*," was translated by GPT-4o and DeepSeek-V3 as شعير مشروب "*barley drink*," omitting the alcoholic context. Likewise, the term *stout*, meaning "*a dark ale*," was transliterated as ستاوت by Gemini 2.0 Flash and DeepSeek-V3—a term not attested in Arabic lexical resources.

[19] We conducted three experiments under different prompt configurations: *Experiment 1*: Each prompt contained a single SL word. *Experiment 2*: Each prompt included 10 SL words, following the same procedure. *Experiment 3*: Each prompt contained 50 SL words, applying the same method. Across all experiments, annotation accuracy for SL words remained consistent. For instance, in all cases in Section 5.4.1, the SL term مهلبية meaning "*a dessert made from milk, starch, and sugar*" was matched to the TL term *pudding*, rather than being classified as a lexical gap.



"I have two lists of words and their definitions. The first list, labeled 'List A,' contains 50 words and their definitions in the source language (SL), as shown below. The second list, in an Excel file labeled 'File B,' contains words and their definitions in the target language (TL). For each SL word in List A, please find its equivalent meaning in the TL from the second list. If no equivalent TL word exists, mark the corresponding SL word as a GAP. Note: If the TL includes a word with an equivalent meaning that is not present in the second list, please provide that TL word along with its definition.
List A contains the following data:

1. *SL_W$_1$: the definition of SL_W$_1$*
2. *SL_W$_2$: the definition of SL_W$_2$*
3. *SL_W$_3$: the definition of SL_W$_3$*

….

…

…

49. *SL_W$_{49}$: the definition of SL_W$_{49}$*
50. *SL_W$_{50}$: the definition of SL_W$_{50.}$"*

FIGURE 8
The template used to prompt each of the three LLMs—GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash.

TABLE 9 Percentage accuracy of validated data collected by GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash models.

Exp. no	Source language	Target language	GPT-4o	DeepSeek-V3	Gemini 2.0 Flash
1	English	Arabic	18	38	14
2	Arabic	English	42	46	38
3	Indonesian	Banjarese	8	10	36
4	Banjarese	Indonesian	40	46	58

5.4.2 Case study: food terminology in Indonesian and Banjarese

This case study investigates lexical diversity in the food domain between Indonesian and Banjarese through two experiments. In the *Indonesian → Banjarese* experiment, we selected 50 culturally specific Indonesian food terms from a dataset of 1,448 entries and compared them against a Banjarese dataset comprising 812 terms. Both datasets were constructed as described in Section 5.3. Using zero-shot prompts, GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash annotated each term by either identifying a Banjarese equivalent or indicating a lexical gap.

The *Banjarese → Indonesian* experiment mirrored this approach in reverse, selecting 50 Banjarese-specific terms and comparing them against the Indonesian dataset. Each language model annotated the terms with corresponding Indonesian equivalents or flagged lexical gaps.

Annotations were reviewed by two native speakers: a university linguistics instructor validated the Banjarese results, while an Indonesian master's student specializing in lexical semantics validated the Indonesian results. Accuracy scores for each model in both translation directions are reported in Table 9, based on expert validation. As in the English–Arabic study, the models frequently made two types of errors: (1) selecting incorrect TL equivalents, and (2) providing literal translations where lexical gaps should have been identified. Examples of such errors include:

- GPT-4o mistranslated the Indonesian word *Beras* "*uncooked rice*" as *Karak* "*hardened rice*."
- DeepSeek-V3 rendered *Papaya* "*a large oval tropical fruit with yellowish flesh*" as *Gandis* "*yellow mangosteen*."
- Gemini 2.0 Flash mistranslated the Banjarese word *Balinjan* "*Tomato*" as *Terung* "*Eggplant*."
- GPT-4o misinterpreted the Banjarese term *Janar* "*Turmeric*" as *Jahe* "*Ginger*."

These findings, together with those from the English–Arabic experiment, underscore the challenges that LLMs face in accurately capturing nuanced or culturally specific lexical meanings—particularly across low-resource language pairs. The bar chart in Figure 9 illustrates the average accuracy across four lexicalisation experiments for three LLMs (GPT-4o, DeepSeek-V3, and Gemini 2.0 Flash), their combined mean ("All LLMs"), and human crowdsourcing. While the LLMs achieve moderate accuracy ranging from 27% to 37%, their combined average of approximately 33% remains substantially lower than the 83.9% accuracy achieved through crowdsourcing. This comparison clearly reveals a pronounced performance gap, indicating that human participants consistently outperform current LLMs in identifying and translating culturally specific lexical items.

6 Perspectives on crowdsourcing

Our crowdsourcing approach offers significant advantages for investigating linguistic diversity. It enables the *systematic* identification of *lexical gaps*, *culture-specific concepts*, and *equivalent word meanings*, thereby providing valuable insights

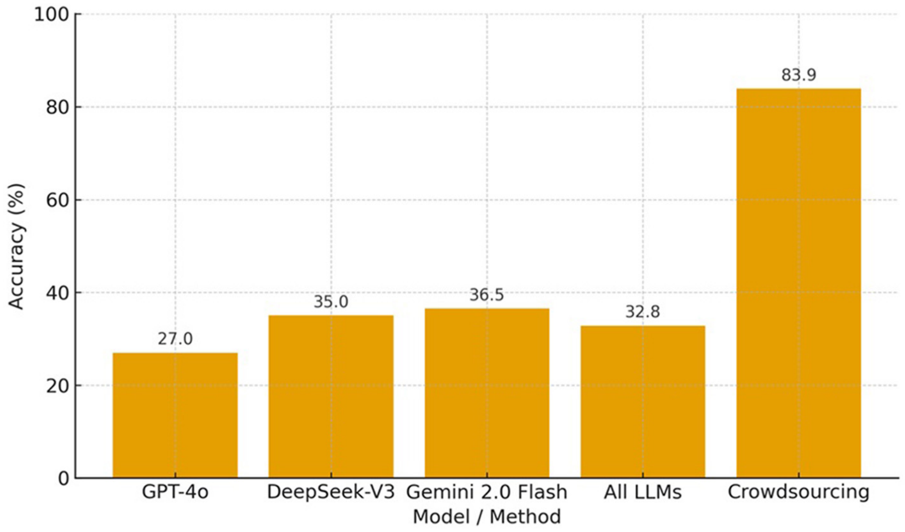


FIGURE 9
Accuracy comparison between LLM and crowdsourced annotations across collected lexical gaps and words.

into language-specific phenomena. The involvement of native speakers ensures contextual precision and captures subtle linguistic and cultural nuances.

This method supports the creation of datasets that are sensitive to lexical diversity and facilitates *bidirectional* exploration between source and target languages, particularly in socially relevant domains such as food. It is both *scalable* and *adaptable*, capable of handling datasets of varying sizes by dividing them into manageable micro-tasks. This task design enhances efficiency and accessibility, even for *non-expert contributors*. For example, in the English–Arabic experiment, each task comprised 35 SL words, while the Indonesian-Banjarese task included 45. Quality control was maintained through real-time validation, *time tracking*, and two *Krippendorff's Alpha–based strategies*: a filtering algorithm during pilot testing to exclude low-quality contributors, and cross-validation of responses using IAA scores across contributor groups.

The *LingoGap* platform streamlines this process with a user-friendly interface, clear instructions, and well-structured workflows. Features such as response time tracking and detailed task logs enhance *transparency* and improve data reliability.

In the long term, the proposed crowdsourcing methodology provides a scalable and sustainable pathway toward more inclusive NLP, enabling the integration of culturally grounded lexical data from underrepresented languages. This integration contributes to richer *multilingual embeddings*, improved *cross-lingual transfer* in large language models, and the sustainable development of linguistic resources across both high- and low-resource languages.

Despite these strengths, the approach presents several challenges. *Timing* and contributor availability significantly affect progress, especially during exam periods, holidays, or personal events. For instance, one task (Task 10 by G_4 in the Arabic-to-English experiment) was delayed by two weeks due to a contributor's maternity leave, while another (Task 7 by G_1 in the Indonesian-to-Banjarese experiment) was postponed because of a family event. These real-world disruptions underscore the importance of flexible task management and careful scheduling, reaffirming earlier findings that the timing of a crowdsourcing campaign is critical, as contributors' availability can be influenced by various external factors (Christoforou et al., 2021).

Recruiting a sufficiently diverse pool of participants remains a challenge, impacting the robustness of evaluation. Additional difficulties arise from the *ambiguity* of certain SL terms—particularly those that fall between well-defined concepts and lexical gaps. For example, the Arabic word مهلبية "*a type of dessert made from milk, starch, and sugar*" and the Banjarese word "rabuk" "*ground meat*" posed classification challenges.

Newly coined and *rare terms* further complicate validation, as they are often under-documented. The approach also struggles to capture *connotative meanings* and affective or metaphorical nuances, which are highly context-dependent and shaped by individual sociocultural backgrounds—dimensions not easily evaluated using standard IAA metrics (Levinson, 2000; Wierzbicka, 1996).

Lastly, *sociolinguistic variation*—including differences based on region, class, age, or gender—is often underrepresented due to demographic biases inherent in many crowdsourcing platforms (Eckert, 2000). Future research should consider employing stratified sampling and sociolinguistically sensitive task design to enhance representativeness.

Together, these perspectives highlight both the potential and the limitations of crowdsourcing in linguistic research. With thoughtful design and continuous refinement, crowdsourcing can serve as a powerful tool for documenting and analyzing lexical diversity across languages.

7 Conclusion

This study addresses the ongoing challenge of linguistic diversity and underrepresentation in multilingual lexical-semantic resources (LSRs), which are essential for many NLP tasks. Current LSRs often reflect English-centric biases and fail to adequately represent lexical gaps, especially in low-resource languages.

To overcome these issues, we introduced a scalable crowdsourcing methodology to systematically collect data on lexical equivalence and gaps across languages. Our approach consists of three key steps: (1) generating input datasets using linguistic resources or language models, (2) gathering responses through the LingoGap crowdsourcing platform, and (3) validating results via IAA and expert review. LingoGap supports the entire process, from task creation to contributor engagement and quality assurance.

We demonstrated the effectiveness of this method through two *large-scale case studies* on food terminology: English–Arabic and Indonesian–Banjarese. These case studies highlight the approach's scalability and its ability to capture lexical diversity in both high- and low-resource languages. Overall, we identified 3,091 lexical gaps and collected 1,735 words across the four languages.

Our findings contribute to building more inclusive and accurate multilingual LSRs, with implications for NLP applications including machine translation, word sense disambiguation, and other NLP tasks.

Future work should extend this methodology to a *broader range of languages and dialects*, particularly underrepresented and typologically diverse ones, to further evaluate its scalability and adaptability. Additionally, exploring *culturally rich semantic domains*—such as body parts (Wierzbicka, 2007), colors (Roberson et al., 2005), visual concepts (Giunchiglia and Bagchi, 2021), emotions, and social values—can provide deeper insights into cross-linguistic and cross-cultural patterns.

Enhancing the data collection process through *real-time expert validation* and incorporating features such as *gamification* and *mobile accessibility* can further improve data quality and increase participation—particularly by engaging broader demographics, including students and the general public.

Another promising direction involves integrating diversity-aware lexical data into LLMs and other NLP systems during both training and evaluation to enhance their cultural and linguistic sensitivity. Such enriched resources can also support a wide range of cross-lingual NLP tasks—including more equitable machine translation, cross-lingual information retrieval, and semantic parsing—as well as interdisciplinary research in education, sociolinguistics, cultural studies (Ono et al., 2023), and linguistic anthropology, particularly for low-resource languages.

In summary, this work offers a practical, scalable framework for advancing linguistic inclusivity in multilingual NLP, promoting more equitable access to language technologies for diverse language communities.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

The studies involving humans were approved by Research Ethics Committee of the University of Trento. The studies were conducted inaccordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HK: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. JO: Conceptualization, Formal analysis, Writing – review & editing. GB: Conceptualization, Formal analysis, Validation, Writing – review & editing. SD: Data curation, Validation, Writing – review & editing. FG: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is additionally funded by the European Union, Grant Agreement No. 101214000 — TRUMAN — HORIZON-CL4-2024-HUMAN-03.

Acknowledgments

We would like to thank the University of Trento and Palestine Technical University—Kadoori for their support. We are also grateful to Prof. Rusma Noortyani from Lambung Mangkurat University, Indonesia, for her assistance in recruiting students for the Indonesian–Banjarese experiment.

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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Author disclaimer

The views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Åke Viberg (1984). *The Verbs of Perception: A Typological Study*. Berlin, Boston: De Gruyter Mouton, 123–162. doi: 10.1515/9783110868555.123

Albala, K. (2011). *Food Cultures of the World Encyclopedia:[4 Volumes]*. New York: Bloomsbury Publishing USA. doi: 10.5040/9798216968511

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J. D., Korobeynikova, M., et al. (2023). Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Armanios, F., and Ergene, B. (2018). *Halal Food: A History*. Oxford: Oxford University Press.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Ling.* 34, 555–596. doi: 10.1162/coli.07-034-R2

Ashley, B., Hollows, J., Jones, S., and Taylor, B. (2004). *Food and Cultural Studies*. London: Routledge. doi: 10.4324/9780203646915

Barbouch, M., Verberne, S., and Verhoef, T. (2021). Wn-bert: Integrating wordnet and bert for lexical semantics in natural language understanding. *Comput. Ling. Netherlands J.* 11, 105–124.

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

Bella, G., Helm, P., Koch, G., and Giunchiglia, F. (2024). "Tackling language modelling bias in support of linguistic diversity," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24* (New York, NY, USA: Association for Computing Machinery), 562–572. doi: 10.1145/3630106.3658925

Benjamin, M., and Radetzky, P. (2014). "Multilingual lexicography with a focus on less-resourced languages: data mining, expert input, crowdsourcing, and gamification," in *9th edition of the Language Resources and Evaluation Conference*.

Biemann, C., and Nygaard, V. (2010). "Crowdsourcing wordnet," in *The 5th International Conference of the Global WordNet Association (GWC-2010)*.

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)*, H. Schuetze, P. Fung, and M. Poesio (Sofia, Bulgaria: Association for Computational Linguistics), 1352–1362.

Brualdi, R. A. (2004). *Introductory Combinatorics*. Noida: Pearson Education India.

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G., Wilie, B., Koto, F., et al. (2023). "NusaCrowd: open source initiative for Indonesian NLP resources," in *Findings of the Association for Computational Linguistics: ACL 2023*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, Canada: Association for Computational Linguistics), 13745–13818. doi: 10.18653/v1/2023.findings-acl.868

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

Christoforou, E., Barlas, P., and Otterbacher, J. (2021). "It's about time: a view of crowdsourced data before and during the pandemic," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY, USA: Association for Computing Machinery). doi: 10.1145/3411764.3445317

Čibej, J., and Arhar Holdt, Š. (2019). "Repel the syntruders! a crowdsourcing cleanup of the thesaurus of modern slovene," in *Proceedings of the ELex 2019 Conference: Electronic lexicography in the 21st century* (Sintra, Portugal).

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., et al. (2023). "Is GPT-3 a good data annotator?," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, Canada: Association for Computational Linguistics), 11173–11195. doi: 10.18653/v1/2023.acl-long.626

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

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Malden, MA: Blackwell Publishers.

El-Haj, M., Kruschwitz, U., and Fox, C. (2015). Creating language resources for under-resourced languages: methodologies, and experiments with arabic. *Lang. Resour. Eval.* 49, 549–580. doi: 10.1007/s10579-014-9274-3

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

Fišer, D., Tavčar, A., and Erjavec, T. (2014). "sloWCrowd: a crowdsourcing tool for lexicographic tasks," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, et al. (Reykjavik, Iceland: European Language Resources Association (ELRA)), 3471–3475.

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). "Crowddb: answering queries with crowdsourcing," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD '11* (New York, NY, USA: Association for Computing Machinery), 61–72. doi: 10.1145/1989323.1989331

Freihat, A. A., Khalilia, H., Bella, G., and Giunchiglia, F. (2024). "Advancing the Arabic WordNet: elevating content quality," in *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, eds. H. Al-Khalifa, K. Darwish, H. Mubarak, M. Ali, and T. Elsayed (Torino, Italia: ELRA and ICCL), 74–83.

Ganbold, A., Chagnaa, A., and Bella, G. (2018). "Using crowd agreement for Wordnet localization," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, eds. N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, et al. (Miyazaki, Japan: European Language Resources Association (ELRA)).

Gantar, P., and Krek, S. (2011). "Slovene lexical database," in *Natural Language Processing, Multilinguality*, 72–80.

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. *Ling. Typol.* 26, 439–487. doi: 10.1515/lingty-2021-2088

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

Giunchiglia, F., Batsuren, K., and Alhakim Freihat, A. (2018). "One world-seven thousand languages (best paper award, third place),"in *International Conference on Computational Linguistics and Intelligent Text Processing* (Springer), 220–235. doi: 10.1007/978-3-031-23793-5_19

Giunchiglia, F., Batsuren, K., Bella, G., et al. (2017). Understanding and exploiting language diversity," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 4009–4017. doi: 10.24963/ijcai.2017/560

Giunchiglia, F., Bella, G., Nair, N. C., Chi, Y., and Xu, H. (2023). Representing interlingual meaning in lexical databases. *Artif. Intell. Rev.* 56, 11053–11069. doi: 10.1007/s10462-023-10427-1

Giunchiglia, F., Jovanovic, M., Huertas-Miguelá nez, M., Batsuren, K., et al. (2015). "Crowdsourcing a large scale multilingual lexico-semantic resource," in *AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)*.

Hasanain, M., Ahmad, F., and Alam, F. (2024). "Large language models for propaganda span annotation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14522–14532. doi: 10.18653/v1/2024.findings-emnlp.850

Kasumba, R., and Neumman, M. (2024). "Practical sentiment analysis for education: the power of student crowdsourcing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 23110–23118. doi: 10.1609/aaai.v38i21.30356

Katsuta, A., and Yamamoto, K. (2020). Lexical simplification by unsupervised machine translation. *Int. J. Asian Lang. Proc.* 30, 2050008. doi: 10.1142/S2717554520500083

Kay, P., and Cook, R. S. (2016). "World color survey," in *Encyclopedia of Color Science and Technology*, ed. M. R. Luo (New York: Springer), 1265–1271. doi: 10.1007/978-1-4419-8071-7_113

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

Khalilia, H., Bella, G., Freihat, A. A., Darma, S., and Giunchiglia, F. (2023). Lexical diversity in kinship across languages and dialects. *Front. Psychol.* 14:1229697. doi: 10.3389/fpsyg.2023.1229697

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*, eds. N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. (Marseille, France: 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. doi: 10.1075/tsl.100

Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T. (2020). "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, eds. D. Scott, N. Bel, and C. Zong (Barcelona, Spain: International Committee on Computational Linguistics), 757–770. doi: 10.18653/v1/2020.coling-main.66

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. Annenberg School for Communication, University of Pennsylvania. Available online at: https://repository.upenn.edu/asc_papers/43 (Accessed March 23, 2024).

Lang, E. (2001). "Spatial dimension terms," in *Language Typology and Language Universals* (Berlin, Boston: De Gruyter Mouton), 1251–1275. doi: 10.1515/9783110194265-028

Lanser, B., Unger, C., and Cimiano, P. (2016). "Crowdsourcing ontology lexicons," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, eds. N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard (Portorož, Slovenia: European Language Resources Association (ELRA), 3477–3484.

Lease, M., and Yilmaz, E. (2012). Crowdsourcing for information retrieval. *SIGIR Forum* 45, 66–75. doi: 10.1145/2093346.2093356

Lehrer, A. (1970). Notes on lexical gaps. *J. Linguist*. 6, 257–261. doi: 10.1017/S0022226700002656

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. London: MIT press. doi: 10.7551/mitpress/5526.001.0001

Lin, Z., Chen, W., Song, Y., and Zhang, Y. (2024). "Prompting few-shot multi-hop question generation via comprehending type-aware semantics," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 3730–3740. doi: 10.18653/v1/2024.findings-naacl.236

Liu, Q., Ihler, A. T., and Steyvers, M. (2013). "Scoring workers in crowdsourcing: how many control questions are enough?," in *Advances in Neural Information Processing Systems*, eds. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Curran Associates, Inc.).

Loureiro, D., and Jorge, A. (2019). Language modelling makes sense: propagating representations through wordnet for full-coverage word sense disambiguation. *arXiv preprint arXiv:1906.10007*.

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

Manerkar, S., Asnani, K., Khorjuvenkar, P. R., Desai, S., and Pawar, J. D. (2022). Konkani wordnet: corpus-based enhancement using crowdsourcing. *Trans. Asian Low-Resour. Lang. inf. Proc*. 21, 1–18. doi: 10.1145/3503156

McCarthy, A. D., Wu, W., Mueller, A., Watson, B., and Yarowsky, D. (2019). Modeling color terminology across thousands of languages. *arXiv preprint arXiv:1910.01531*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology* 9, 165–208. doi: 10.2307/3772782

Nair, N. C. (2022). A crowdsourcing methodology for improving the Malayalam Wordnet. *SSRN Electron. J*. 4064783. doi: 10.2139/ssrn.4064783

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* (Institute of Digital Enhancement of Cognitive Processing, Waseda University), 255–264.

Ono, M., Soga, T., Kikuchi, M., and Tanabe, T. (2023). "Consideration of language learning service with visualized vocabulary map derived from wordnet," in *2023 8th International Conference on Business and Industrial Research (ICBIR)* (IEEE), 1194–1198. doi: 10.1109/ICBIR57571.2023.10147578

Parent, G., and Eskenazi, M. (2010). "Clustering dictionary definitions using Amazon Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, eds. C. Callison-Burch, and M. Dredze (Los Angeles: Association for Computational Linguistics), 21–29.

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

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

Post, M., Callison-Burch, C., and Osborne, M. (2012). "Constructing parallel corpora for six Indian languages via crowdsourcing," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, eds. C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia (Montréal, Canada: Association for Computational Linguistics), 401–409.

Powers, D. M. W. (2012). "The problem with kappa," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ed. W. Daelemans (Avignon, France: Association for Computational Linguistics), 345–355.

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

Robinson, J., Rosenzweig, C., Moss, A. J., and Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PLoS ONE* 14:e0226394. doi: 10.1371/journal.pone.0226394

Sayeed, H. M., Mohanty, T., and Sparks, T. D. (2024). Annotating materials science text: a semi-automated approach for crafting outputs with gemini pro. *Integr. Mater. Manuf. Innov*. 13, 445–452. doi: 10.1007/s40192-024-00356-4

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

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Penyusun Kamus Pusat Bahasa (2008). *Kamus Bahasa Indonesia*. Pusat Bahasa Departemen Pendidikan Nasional.

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

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

Warrens, M. J. (2011). Cohen□s kappa is a weighted average. *Stat. Methodol*. 8, 473–484. doi: 10.1016/j.stamet.2011.06.002

Welbl, J., Liu, N. F., and Gardner, M. (2017). Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*. doi: 10.18653/v1/W17-4413

Wierzbicka, A. (1996). *Semantics: Primes and Universals: Primes and Universals*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198700029.001.0001

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

Wijesiri, I., Gallage, M., Gunathilaka, B., Lakjeewa, M., Wimalasuriya, D., Dias, G., et al. (2014). "Building a WordNet for sinhala," in *Proceedings of the Seventh Global Wordnet Conference*, eds. H. Orav, C. Fellbaum, and P. Vossen (Tartu, Estonia: University of Tartu Press), 100–108. doi: 10.18653/v1/W14-0114