UNIVERSITÀ
DI TRENTO

Department of Information Engineering and Computer Science

Doctoral Thesis

Developing Language Resources: A Lexical-Diversity-Centric Approach

Author:
Hadi Khalilia

Supervisors:
Prof. Fausto Giunchiglia
Prof. Gábor Bella

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

April 2, 2025

Contents

List of Figures

List of Tables

Acronyms

ACQ Attention Check Question

AI Artificial Intelligence

AWN Arabic WordNet

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pre-Trained Transformer

GUI Graphical User Interface

IAA Inter-Annotator Agreement

LLM Large Language Model

LSRs Lexical-Semantic Resources

MSA Modern Standard Arabic

MTurk Amazon Mechanical Turk

MT Machine Translation

NLP Natural Language Processing

NLU Natural Language Understanding

POS Part of Speech

PWN Princeton WordNet

SF Semantic Field

SL Source Language

SUMO Suggested Upper Merged Ontology

TL Target Language

UKC Universal Knowledge Core

UX/UI User Experience/User Interface

WSD Word Sense Disambiguation

Abstract

Languages describe the world in diverse ways, a phenomenon known as linguistic diversity, which is also reflected in their lexical-semantic resources (LSRs). These resources, such as online lexicons and WordNets, are essential for natural language processing (NLP) applications. However, in many languages, this diversity often results in pervasive quality issues, including inaccuracies, incompleteness, and, notably, a bias towards the English language and Anglo-Saxon culture. This bias is evident in the omission of concepts unique to specific languages or cultures, the inclusion of foreign (Anglo-Saxon) concepts, and the lack of explicit indications for untranslatability. The latter, referred to as cross-lingual lexical gaps, occurs when a term has no equivalent in another language.

The development of diversity-aware multilingual lexical resources faces significant challenges, particularly in addressing cross-lingual lexical gaps and capturing linguistic diversity in low-resource languages. Current approaches often lack systematic methods to identify lexical untranslatability, leading to resources that inadequately represent the linguistic and cultural richness of diverse languages. Low-resource languages are especially underserved due to the lack of comprehensive lexical databases and the limitations of automated methods in capturing linguistic nuances. Existing practices, which are often unidirectional and rely on English as a pivot language, exacerbate English bias and distort the representation of non-English linguistic concepts. Furthermore, expert-driven methods typically focus narrowly on advanced-level lexical gaps in a limited subset of languages, leaving much of the world's linguistic diversity unaddressed. These challenges impede the creation of high-quality, inclusive datasets essential for advancing NLP applications, particularly for underrepresented languages.

This thesis presents a systematic hybrid approach for generating diversity-aware, multilingual lexical-semantic resources by integrating an expert-driven method with a novel crowdsourcing methodology. The expert-driven method involves language specialists in a structured process to identify lexical gaps, encompassing contribution collection, validation, and concept-level verification. The crowdsourcing methodology complements this by facilitating the bidirectional exploration of lexical gaps between language pairs without relying on English as an intermediary, leveraging contributions from native speakers.

Based on these considerations, the main contributions of this research are as follows. *(i)* a systematic hybrid approach consisting of an expert-driven method and a novel crowdsourcing methodology for collecting data on linguistic diversity. *(ii)* development of LingoGap, a crowdsourcing platform specifically designed to gather linguistic diversity data from ordinary native speakers. *(iii)* validation of the expert-driven method through five

case studies, including: (a) three large-scale studies conducted on kinship terminology across seven Arabic dialects, three Indonesian languages, and ten languages from various language families. (b) a study focusing on basic-level categories in six languages: Arabic, Turkish, Persian, Indonesian, Banjarese, and Javanese. (c) a study aimed at enhancing the Arabic WordNet to address linguistic diversity by restructuring its content and incorporating lexical gaps and Arabic-specific concepts. *(iv)* validation of the crowdsourcing methodology using LingoGap through eight large-scale experiments, including: (a) two studies on food-related terminology across English-Arabic and Indonesian-Banjarese language pairs. (b) six studies on kinship-related terms involving the following language pairs: English-Persian, Arabic-Persian, English-Turkish, as well as comparisons between language-independent kinship concepts and Turkish, Persian, and Maba. *(v)* publication of the resulting data, which consists of 7,273 lexical gaps, 8,497 equivalent words, and 9,576 high-quality Arabic synsets—sets of synonymous words—as open, computer-processable datasets, along with their integration into the Universal Knowledge Core multilingual database. *(vi)* development of high-quality, diversity-aware lexicons includes the following: Arabic WordNet (V3), the first diversity-aware lexicon for Standard Arabic; ArabicUKC, a multilingual lexicon for Arabic dialects; and IndonesianUKC, a multilingual lexicon for Indonesian languages. *(vii)* evaluation of the proposed methodology in comparison with large language models (LLMs), focusing on data annotation via crowdsourcing versus LLM agents in two case studies. Results demonstrate that crowdsourcing significantly outperforms LLMs in identifying language- and culture-specific concepts. *(viii)* introduction of a machine translation evaluation methodology using the produced diversity-aware datasets.

Keywords: Language resources, Multilingual Lexicon, Language Diversity, Crowdsourcing, Expert-Sourcing, Linguistic Gap, Lexical Typology.

Acknowledgements

Completing this PhD has been a journey of challenges, discoveries, perseverance, and personal growth. It has been shaped by the guidance, encouragement, and support of many individuals, to whom I am deeply grateful. First and foremost, I would like to express my heartfelt gratitude to my advisors for their unwavering support and invaluable guidance, which have been integral to the success of this research. I am profoundly thankful to my supervisor, Prof. Fausto Giunchiglia, for his mentorship, vision, and encouragement throughout this journey. I am also grateful to my co-supervisor, Prof. Gábor Bella, whose expertise, guidance, and support—both academically and during my research period abroad—have been invaluable. Furthermore, I wish to acknowledge the DECIDE research group at IMT Atlantique, France, for providing the opportunity and support that facilitated my research period abroad.

I extend my sincere appreciation to Dr. Abed Alhakim Freihat for his insightful feedback and guidance on the expert-sourcing aspects of my work, and to Dr. Jahna Otterbacher for her invaluable support and advice on the crowdsourcing part of this research. I am deeply grateful to my thesis reviewers, Prof. Enrico Giunchiglia and Dr. Wajdi Zaghouani, for their constructive feedback and thoughtful suggestions, which have significantly enhanced the quality of this thesis. I also thank the committee members — Prof. Enrico Giunchiglia, Dr. Wajdi Zaghouani, and Prof. Armando Tacchella — for their time and insightful comments during the defense.

Special thanks go to Prof. Rusma Noortyani and Shandy Darma for their exceptional coordination of the experiments involving the Indonesian, Javanese, and Banjarese languages; to Kubra Korkmaz for her dedicated coordination of the Turkish experiments; to Arya Torabi for his efforts in coordinating the Persian experiments; and to Mahamat Choueb for coordinating the experiments on the Maba language. I would also like to express my gratitude to Yamini Chandrashekar, Ali Hamza, Alessio Zamboni, Andrea Bontempelli, and all the members of KnowDive for their insightful discussions, continuous support, and for fostering a welcoming and inspiring working environment.

Lastly, I am profoundly thankful to my family and friends for their unconditional patience and encouragement throughout this journey. Their steadfast support has been my greatest source of strength and motivation.

1 Introduction

Contents

This chapter provides an overview of the work of this thesis. We start by describing the motivation and the context that guided the development of the methodologies and tools presented in the next chapters. Then, the thesis's contributions in tackling the relevant problems are outlined. Finally, we list the publications on which the thesis is based and also a brief summary of each chapter.

1.1 Motivation

In recent years, the rise of deep learning techniques, especially large language models, has captured much attention within the natural language processing NLP community. Yet, *lexical-semantic resources* (LSRs) remain fundamental pillars in applications such as machine translation, word sense disambiguation, and information retrieval [76, 34, 100, 17]. However, many existing LSRs, like WordNets, exhibit significant language biases—particularly toward English [24, 67]. This bias stems largely from the pervasive use of the Princeton WordNet (PWN) [106] as the foundation for translations, often leading to incomplete or inaccurate lexical resources for non-English languages [81, 82]. Consequently, language- or culture-specific concepts, especially those lacking direct English equivalents, are frequently underrepresented. This is not only leading to an incomplete or skewed understanding of linguistic diversity [65] but also limits the ability of NLP systems to embrace the full spectrum of human language [83].

Language- and culture-specific concepts and *lexical untranslatability* pose significant challenges in NLP. *Language- and culture-specific concepts* are

words or expressions that exist in one language or culture but lack exact equivalents in others. They reflect the unique experiences, values, or traditions of a particular group of people, shaped by their linguistic and cultural practices. For example, in the Italian Alps, the word *malga* refers to "*a traditional mountain restaurant*". This term is deeply rooted in Italian culture and has no direct equivalent outside the Alpine region [72].

Such concepts and differences between linguistic communities lead to *lexical untranslatability*, also known as *cross-lingual lexical gaps*. This phenomenon occurs when a word or expression exists in one language but has no direct equivalent in another. For example, the Arabic word خالة, meaning "*the sister of your mother*", has no direct counterpart in English—just as the English term *sibling* lacks an exact equivalent in Arabic or languages such as Italian and Japanese [84]. These lexical gaps can significantly impact the accuracy of machine translations and the quality of semantic resources. For instance, when translating the English sentence *his cousin gave birth to twins*, Google Translate produces the Arabic output أنجب ابن عمه توأما (a'njaba ibna a'mihi tawaman), meaning "*his father's brother's son gave birth to a twin*". Though syntactically correct, this translation introduces the unintended meaning of a male giving birth, highlighting a lexical gap—specifically, the absence of an equivalent Arabic term for *cousin* [80]. This issue is further exacerbated in the context of low-resource languages, which are often neglected in existing multilingual lexical resources, thus widening the gap in linguistic representation [83].

Addressing such instances of lexical untranslatability requires a systematic approach to developing diversity-aware datasets. By ensuring the comprehensive representation of linguistic and cultural nuances, such datasets can provide a more robust foundation for cross-lingual NLP applications.

1.2 Problem

Developing diversity-aware datasets for multilingual lexical resources remains a significant and persistent challenge within NLP. Despite the importance of addressing cross-lingual lexical gaps, *no comprehensive method* has yet been introduced to *systematically identify these gaps*. Consequently, large-scale resources addressing lexical untranslatability across languages are scarce, with only a handful of limited examples currently available. Although some progress has been made toward systematization in specific domains—such as kinship and color terminology within ethnography, linguistics, and cognitive science, as demonstrated in our kinship experiments (Khalilia et al. [80] and Khishigsuren et al. [84])—the majority of fields remain underexplored.

Furthermore, *few efforts* have been made to address the exploration of language diversity in *low-resource languages*, which frequently falls outside

the focus of mainstream NLP research. Low-resource languages lack comprehensive lexical databases, and automatic approaches struggle to accurately capture their linguistic complexity. This results in a lack of quality data to support machine translation and other NLP applications in these languages, further widening the digital divide between high-resource and low-resource languages [80, 83]. The limited focus on low-resource languages hinders the development of diversity-aware datasets that could capture the full range of linguistic and cultural variations across different language communities.

Another significant issue is that many approaches explore language diversity in *a unidirectional manner*, typically from the source language (often English) to the target language. This practice results in a narrow understanding of language diversity, as it fails to capture gaps that may exist when translating in the reverse direction [83]. Moreover, the reliance on English as a *pivot* language in translation processes leads to what is known as *English bias*, where certain linguistic or cultural concepts that are unique to non-English languages are not adequately represented [24]. This bias affects the quality of the generated datasets, as it skews the representation of lexical diversity toward English-centric frameworks [72, 65].

These limitations are further compounded by the insufficient quality of automatically generated datasets. *Automatic methods* often fail to capture the linguistic nuances and cultural specificities inherent in non-English languages, particularly in domains rich in diversity, such as kinship or food-related terminology, as we observed in our research (Khishigsuren et al. Khishigsuren et al. [84]) while exploring specific kinship concepts in Arabic, Malayalam, and Mongolian. Consequently, the datasets produced by these methods lack the depth and accuracy needed for effective NLP applications, especially in the context of low-resource languages [19, 20].

Finally, existing *expert-driven approaches* for translating source datasets, primarily English resources, such as PWN—into target languages have largely centered on identifying lexical gaps at an *advanced level* within the target language, specifically where experts possess domain-specific knowledge. Moreover, these methods have been applied to only *a narrow subset of natural languages*, despite the global linguistic landscape encompassing over seven thousand languages.

These challenges highlight the need for a more innovative, efficient, and systematic approach to identifying and documenting a wider spectrum of lexical gaps. Such an approach would be instrumental in creating high-quality, diversity-aware datasets, thereby advancing the effectiveness of cross-lingual NLP applications.

1.3 Solution

This research presents a systematic, *hybrid approach* for creating diversity-aware datasets. With the aim of increasing the collection of lexical gaps, this approach consists of two methods: an *expert-driven method*—a top-down method that draws on the expertise of language specialists, thus gathering information from diverse language users, and *a novel crowdsourcing method*—a bottom-up method that relies on input from the general public, specifically native speakers. This flexible approach allows each method to be used independently or in combination to identify instances of lexical untranslatability.

Firstly, the *expert-driven method*, widely used for creating small-scale datasets of lexical gaps in domains such as kinship and color, is extended in this study to cover a wider range of languages, enabling the generation of large-scale, diversity-aware datasets. This approach employs two human translators working from a source language (SL) to a target language (TL) to systematically identify lexical gaps, followed by validation from a target language specialist. Inputs for this method include an SL word list with definitions and a linguistic resource, such as a dictionary. The process consists of four main steps:

1. *Contribution task generation:* This step involves collecting SL data from the Universal Knowledge Core (UKC), a high-quality, diversity-aware multilingual lexicon that encompasses millions of words and meanings from over 2,300 languages [65]. The data is then prepared in a spreadsheet file to be provided for translation into the TL. Each SL entry comprises a tuple (word, gloss), representing a term and its definition. Additionally, the data is formalized based on the UKC concept hierarchy.

2. *Contribution collection*: The actual contribution effort is carried out by a native speaker in the TL language. For each SL word in the spreadsheet, the translator determines whether it has an equivalent meaning in the TL or constitutes a lexical gap.

3. *Lexicon-level validation*: Provided words and gaps are evaluated and corrected by a language expert.

4. *Concept-level validation*: New concepts and unclear contributions (i.e., words on the borderline) are verified by a lexico-semantic expert.

We applied our expert-driven method to produce diversity-aware datasets through two case studies. The first is a *monolingual* experiment in which a list of language-independent concepts from a specific semantic field is lexicalized into a target language. This study focused on kinship terminology and basic-level categories, as detailed below:

1. **Exploring linguistic diversity across languages and dialects** [80]: We employed the expert-driven method to construct diversity-aware lexical resources across local languages and dialects, validated through two comprehensive case studies focused on kinship terms in seven Arabic dialects and three Indonesian languages. We systematically compiled 223 kinship terms and identified 1,619 lexical gaps, including 22 previously undocumented (new) kinship concepts in the languages represented in the UKC. This work underscores the profound linguistic diversity that exists even among dialects of the same language and enhances current lexical resources by incorporating these findings into the UKC.

2. **Translating basic-level Categories:** Using the expert-driven method, we translated 965 concepts from basic-level categories into six languages: Arabic, Turkish, Persian, Indonesian, Javanese, and Banjarese. In total, we collected 28 gaps and 940 words in Arabic, 65 gaps and 903 words in Persian, 15 gaps and 953 words in Turkish, 20 gaps and 948 words in Indonesian, 19 gaps and 949 words in Javanese, and 12 gaps and 956 words in Banjarese. Altogether, this process resulted in the collection of 159 gaps and 5,649 words.

3. **Enrichment of the KinDiv lexical resource**[84]: KinDiv[1] is a large-scale lexico-semantic resource that includes 198 kinship concepts, 1,911 words, and 37,370 identified gaps across 699 languages. This resource was collaboratively developed by the National University of Mongolia and the University of Trento. Its construction employed a hybrid approach, utilizing three main sources of evidence for lexicalization: Murdock's lexicalization patterns [111], entries from Wiktionary, and direct input from native speakers. The latter represents our primary contribution of this research, applying the expert-driven methodology to enrich the resource with lexical gaps and kinship terms from 10 languages: Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish. This effort resulted in the identification of 1,503 lexical gaps and the addition of 396 words contributed by native speakers of these languages.

The second case study involves a *bilingual* experiment translating a list of English words into a target language. This study focused on English and Arabic, resulting in the development of a high-quality, diversity-aware Arabic WordNet. By employing the expert-driven methodology, we achieved significant advancements in Arabic WordNet [51], addressing key challenges related to correctness, completeness, and linguistic diversity. Our efforts improved over 58% of synsets by supplementing missing components and

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

correcting errors. Additionally, we introduced innovative elements such as *phrasets*—word combinations used to express synset meanings when no direct equivalent lemmas exist—and *lexical gaps*, which explicitly represent untranslatable concepts, thereby enriching the resource's capacity to capture linguistic diversity. This work culminated in the release of *AWN V3* [60], a substantially more accurate, comprehensive, and accessible version of Arabic WordNet, optimized for both human use and NLP applications.

Secondly, the *novel crowdsourcing methodology* that enables the exploration of lexical diversity *bidirectionally* between any two languages—from the source language (SL) to the target language (TL) and vice versa. Unlike traditional methods, this approach *does not rely on English* as an intermediate language (e.g., a pivot), thereby avoiding the biases that often distort linguistic diversity. By capturing lexical gaps and equivalent terms across languages, including low-resource ones, this scalable method offers a viable alternative to conventional approaches. It generates high-quality, diversity-rich data and facilitates applicability across language pairs, contingent upon the availability of bilingual native speakers as demonstrated in our study Khalilia et al. [83].

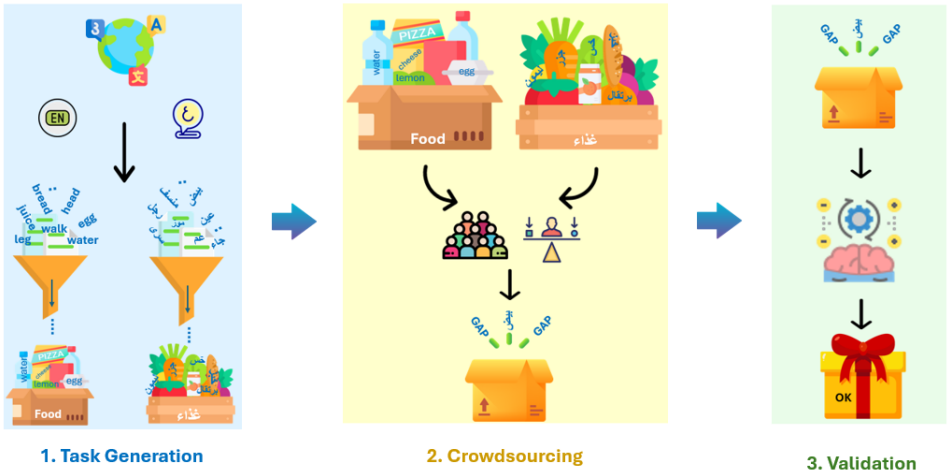


Figure 1.1: General Overview for the Crowdsourcing Methodology.

The initial *inputs* to our crowdsourcing methodology are an *ordered source–target language pair* and a specified *semantic field*. An overview of the crowdsourcing method is depicted in Figure 1.1. It includes three main steps:

1. *Task generation:* A semi-automated process that generates two lists of lexical entries, one for each language, where each entry is a tuple (word, gloss), i.e., a term and its definition.

2. *Crowdsourcing:* Native speakers compare a lexical entry from the SL with all corresponding TL entries, determining translation equivalence or identifying lexical gaps.

3. *Validation* Lexical entries and gaps are evaluated automatically, followed by expert-based verification.

At the core of this methodology is the use of crowdsourcing platforms, where native speakers engage in micro-tasks through the *LingoGap* developed tool. This method effectively explores cross-lingual lexical gaps in culturally diverse domains, such as food [11], kinship [84], and body parts [159]. Furthermore, it can also be applied to investigate lexical gaps in more common domains with low linguistic diversity, such as basic-level categories [131].

To ensure the collection of *high-quality crowdsourced data*, we implemented a three-phase quality control framework. In the *pre-crowdsourcing* phase, we selected qualified workers through a proficiency test and applied an algorithmic filter based on Krippendorff's Alpha to exclude low-quality annotators. During the *task execution*, real-time quality control measures were employed, including attention check questions and completion time monitoring, to detect inattentive or inconsistent contributors. In the *post-crowdsourcing* phase, we validated the collected responses through automatic inter-annotator agreement analysis and expert review of borderline cases, ensuring that only reliable data were retained.

This methodology was validated through two case studies. The first involved five experiments across *language pairs*, including the following:

1. The English-Arabic experiment on the food terminology.

2. The Indonesian-Banjarese experiment on the food terminology.

3. The Arabic-Persian experiment on the kinship terminology.

4. The English-Persian experiment on the kinship terminology.

5. The English-Turkish experiment on the kinship terminology.

The second case study involves *monolingual experiments*, in each experiment, a list of language-independent concepts from a specific semantic field is compared to a list of words from the same field in a single language. This study included three experiments focused on kinship concepts, as follows:

1. The first experiment focuses on comparing kinship concepts with Persian kinship terms.

2. The second experiment focuses on comparing kinship concepts with Turkish kinship terms.

3. The third experiment focuses on comparing kinship concepts with Maba (a language spoken in eastern Chad by approximately 300,000 speakers) kinship terms.

A total of 3,690 lexical gaps and 1,872 equivalent words were identified using the crowdsourcing methodology. This method proved highly effective for enriching lexicons of low-resource languages, significantly outperforming large language models in capturing *language- and culture-specific concepts*. In addition to the hybrid approach and the contributions mentioned earlier, this work provides the following additional insights:

1. **The quality of lexical-semantic resources** [81, 82]: We conducted an extensive survey and introduced a systematic framework for assessing the quality of lexical-semantic resources, focusing on two critical dimensions: the accuracy and completeness of a synset (a set of synonymous words). Central to our approach is a rigorous evaluation process that examines key synset components—lemmas, glosses (natural language definitions encapsulating synset meanings), and illustrative examples—along with the semantic relationships connecting synsets. This framework addresses pervasive issues such as incorrect or missing senses, and incorrect or incomplete relations, which frequently compromise the effectiveness of these resources in NLP applications. Our contributions can be leveraged to enhance the precision of lexical-semantic resources, thus enabling more accurate and efficient deployment in language processing tasks.

2. **UKC enrichment:** We integrated the linguistic knowledge gained from the efforts described above into the UKC by adding lexical gaps, words, and new concepts. Specifically, 7,273 lexical gaps, 8,497 equivalent words, and 9,576 Arabic synsets (comprising 14,535 words, 9,322 new glosses, and 12,204 new example sentences) were incorporated into the UKC lexicons.

3. **Developing Arabic and Indonesian UKCs**: We developed two instances of the UKC. The first is the Arabic UKC[2], which includes lexicons from Standard Arabic as well as various Arabic dialects. The second is the Indonesian UKC[3], a dedicated instance of the UKC for Indonesian languages. The Arabic data collected from the kinship experiments described above has been integrated into the Arabic UKC. Similarly, the Indonesian kinship data has been incorporated into the Indonesian UKC.

[2] https://arabic.ukc.datascientia.eu
[3] https://indonesia.ukc.datascientia.eu

4. **Developing a methodology for machine translation evaluation:** We have developed a methodology for evaluating machine translation systems, utilizing diversity-aware datasets integrated into the UKC. This methodology was validated through the evaluation of three MT systems: Google Translate, GPT-4, and DeepL. The evaluation employed a corpus of kinship sentences created in Arabic, Turkish, and Indonesian.

1.4 List of publications

This thesis is based on the following publications:

- **Hadi Khalilia**, Jahna Otterbacher, Gábor Bella, Rusma Noortyani, Shandy Darma and Fausto Giunchiglia. 2025 "Crowdsourcing Lexical Diversity". *Language Resources and Evaluation Journal*. **Under review**.

  https://arxiv.org/abs/2410.23133

- Abed Alhakim Freihat, **Hadi Khalilia**, Gábor Bella, and Fausto Giunchiglia. 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 co-located LREC-COLING 2024*.

  https://aclanthology.org/2024.osact-1.9/

- **Hadi Khalilia**, Gábor Bella, Abed Alhakim Freihat, Shandy Darma and Fausto Giunchiglia. 2023. "Lexical diversity in kinship across languages and dialects". *Frontiers in Psychology, 14*.

  https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1229697/full

- Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, **Hadi Khalilia**, Yamini Chandrashekar, Fausto Giunchiglia. 2022. "Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship". *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

  https://aclanthology.org/2022.lrec-1.299.pdf

- **Hadi Khalilia**, Abed Alhakim Freihat and Fausto Giunchiglia. 2021. "The Dimensions of Lexical Semantic Resource Quality". *In Proceedings of the Second International Workshop on NLP Solutions for*

Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021.

https://aclanthology.org/2021.nsurl-1.3.pdf

- **Hadi Khalilia**, Abed Alhakim Freihat and Fausto Giunchiglia. 2021. "The Quality of Lexical Semantic Resources: A Survey". *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021).*

  https://aclanthology.org/2021.icnlsp-1.14.pdf

- Gábor Bella, **Hadi Khalilia** and Fausto Giunchiglia. "A Lexical Diversity Challenge for Machine Translation". **Under submission**.

1.5 Outline

This section provides a roadmap of the thesis, outlining its main components and objectives. The first part, *Lexical-Semantic Resources and Linguistic Diversity*, establishes the theoretical foundation of the thesis by examining the role of lexical-semantic resources (LSRs) in capturing linguistic diversity. It discusses the challenges of untranslatability, the limitations of existing LSRs, and the need for diversity-aware resources to represent untranslatability and language-specific concepts (Chapters 2 and 3). The second part, *Expert-Sourcing Lexical Diversity*, introduces a structured, expert-driven methodology for documenting lexical diversity (Chapter 4). It details how experts systematically identify lexical gaps and validate contributions across multiple languages. Case studies include kinship terminology, basic-level categories, and enhancements to the Arabic WordNet (Chapter 5). The third part, *Crowdsourcing Lexical Diversity*, presents a complementary crowdsourcing approach to collecting lexical diversity and enhancing resource inclusivity. It outlines the methodology, supported by the LingoGap platform, which engages native speakers in identifying lexical gaps and equivalent terms across language pairs without relying on English as a pivot, thereby mitigating language bias (Chapters 6 to 9). Finally, the fourth part, *Resulting Datasets and Applications*, describes the datasets generated through expert-sourcing and crowdsourcing, including their integration into the Universal Knowledge Core to improve linguistic diversity representation in underrepresented languages. It highlights applications in NLP, such as machine translation and diversity-aware lexicon development, demonstrating the broader impact of this work. The rest of this work is structured as follows:

Chapter 2 discusses the foundational role of LSRs in NLP, highlighting monolingual and multilingual resources, their applications, and the challenges in ensuring quality and inclusiveness across languages.

Chapter 3 explores the implications of linguistic diversity for technology, including issues of untranslatability, and proposes frameworks to better represent diversity in lexical databases.

Chapter 4 describes a structured, expert-driven process for systematically identifying lexical gaps, validating contributions, and enriching multilingual resources through expert input.

Chapter 5 presents case studies that apply the expert-driven methodology to basic-level categories and kinship terminology across multiple languages. These efforts enrich resources such as the UKC by addressing lexical gaps, identifying equivalent term meanings, and documenting new linguistic concepts, while also contributing to the enhancement of the Arabic WordNet.

Chapter 6 explores crowdsourcing as a versatile approach for data collection and annotation in NLP and computational linguistics. It provides an overview of crowdsourcing, defining its concept, outlining its applications and system architecture, and highlighting its pivotal role in fostering diversity-aware multilingual lexicons.

Chapter 7 introduces a crowdsourcing methodology, describing the step-by-step process for designing tasks, collecting contributions, and validating crowdsourced data to develop diversity-aware lexical resources.

Chapter 8 describes the development and evaluation of the LingoGap platform—a tool designed to support the crowdsourcing of lexical diversity data—focusing on its user experience and integration within the research community.

Chapter 9 details a series of experiments conducted using the proposed crowdsourcing methodology to document lexical diversity, focusing on domains like food and kinship across various language pairs.

Chapter 10 summarizes the datasets generated through expert-driven and crowdsourcing methods, highlighting their contribution to documenting lexical gaps, equivalent terms, and linguistic insights.

Chapter 11 demonstrates the application of the developed datasets in integrating with multilingual databases, creating diversity-aware lexicons, enhancing and evaluating machine translation.

Chapter 12 summarizes the thesis's contributions, discusses the benefits and limitations of the research, examines its ethical implications, and outlines future directions for advancing the representation of linguistic diversity.

Part I

Lexical-Semantic Resources and Linguistic Diversity

2

Lexical-Semantic Resources

Contents

Lexical semantic resources (LSRs) are specialized lexical databases designed to organize and represent the relationships between words and their meanings through both lexical and semantic relations. LSRs form the backbone of many applications in NLP and computational linguistics.

In this chapter, we first explore a monolingual lexical resource in Section 2.1, which maps words within a specific language. Next, we describe multilingual lexical resources in Section 2.2, highlighting semantic structures across different languages. Section 2.3 examines the applications of LSRs such as machine translation, natural language understanding, word sense disambiguation, and semantic analysis. Section 2.4 discusses the dimensions of LSR quality, including content accuracy and coverage, and analyzes key factors influencing quality, such as *linguistic diversity*. Finally, Section 2.5 concludes the chapter.

2.1 Monolingual lexical resource

A monolingual LSR is a lexicon that provides information about words and their meanings within a single language. It can take various forms, including traditional dictionaries, thesauruses, and digital databases used in computational linguistics and NLP, for example, Oxford English Dictionary (OED)[1], which offers comprehensive information about English words, their etymology, definitions, and usage. Examples of monolingual lexical databases include wordnets such as Princeton WordNet (PWN) [106, 54] for English,

[1] https://www.oed.com

which is the most prominent example in the NLP community, Polish WordNet [121] for Polish and Arabic WordNet [51] for Standard Arabic. Wordnet groups words into sets of synonyms called *synsets*.

A synset, or synonym set, is a collection of lemmas, each representing the dictionary or canonical form of a word. Each synset includes three primary components: lemmas, a gloss, and example sentences. The lemma is the headword or base form of the word, often used to standardize word variations (e.g., "*run*", rather than "*running*" or "*ran*"). The gloss provides a concise, natural language definition, offering readers or systems a clear sense of the synset's meaning. Example sentences further illustrate the context and nuances of the lemma's meaning within the synset, grounding the gloss in practical usage [106]. Consider the following synset example, drawn from PWN:

Example 1 *"(n) cake: baked goods made from or based on a mixture of flour, sugar, eggs, and fat; she baked a delicious cake for the celebration."*

In this example, "*cake*" serves as the lemma, representing the headword of the synset. The gloss, "*baked goods made from or based on a mixture of flour, sugar, eggs, and fat*", provides a descriptive definition of cake in its edible sense. The example sentence, "*she baked a delicious cake for the celebration*", offers contextual usage, illustrating the concept and use of the word in a sentence. This structure supports both human and machine understanding by linking definitions to real-world examples.

LSRs enhance comprehension and interconnectivity among synsets by leveraging various types of relationships. These relationships are generally classified into two primary categories: *lexical relations* and *semantic relations*. A *lexical relation*, which organizes relationships between different meanings of words, aiding in language processing by highlighting interactions of meaning and opposition. This type of relation is particularly prominent among adjectives and verbs. Some examples are: (1) *antonymy* which connects Opposite meanings (e.g., "*hot*" vs. "*cold*"); (2) *synonymy* which links similar meanings (e.g., "*big*" vs. "*large*"); (3) *homonymy*, which connect unrelated meanings with same form (e.g., "*bat*" as an animal or sports equipment).

A *semantic relation* structures the relationships between synsets, organizing them based on conceptual and hierarchical links. Such relationships are more commonly found between nouns or verbs, where the connection goes beyond a mere opposition or similarity and involves categorical or part-whole relationships. For example, the synset "*cake*" has several semantic relationships with other synsets, showcasing a network of associations that help define its place within broader conceptual structures, as illustrated in Figure 2.1. Some examples of these relations are:

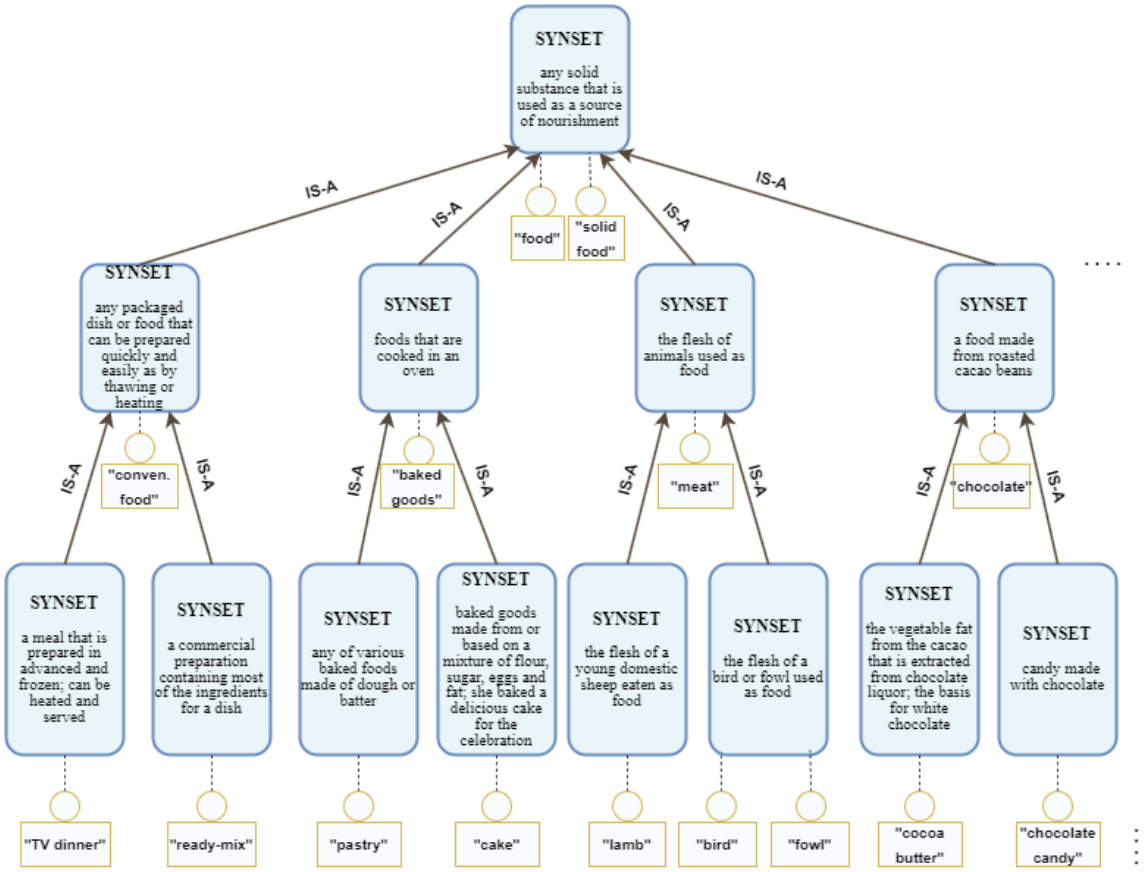


Figure 2.1: Mapping of word meanings in the PWN.

1. *Hypernyms*, which denote broader categories under which more specific concepts fall. For instance, "*baked goods*" is a hypernym for "*cake*".

2. *Hyponyms*, in contrast to hypernyms, hyponyms represent more specific terms within a broader category. For example, "*cheesecake*" is a hyponym of "*cake*".

3. *Meronyms*, which describe part-whole relationships, where a component or ingredient forms part of a larger concept. For example, "*flour*" is a meronym of "*cake*".

4. *Holonyms*, which conversely describe whole-part relationships, where a larger concept encompasses smaller components. For example, "*cake*" can be a holonym of "*birthday party*".

2.2 Multilingual lexical resource

A multilingual lexical resource is a structured collection of words, phrases, and linguistic information across multiple languages. It is designed to facilitate understanding, translation, and language processing by offering meanings and linguistic details such as pronunciation, grammar, and usage. Examples include multilingual dictionaries, thesauri, and linguistic databases. A multilingual lexicon encompasses both lexical and semantic relationships between languages. These resources enable NLP systems to interpret and link meanings across languages, supporting cross-lingual applications such as machine translation, information retrieval, and sentiment analysis.

The process of cross-lingual lexical mappings in multilingual LSRs represents a major challenge in the development of such resources, particularly in addressing lexical equivalence among word meanings in two (or more) languages. Linguists understand lexical equivalence as a complex and multidimensional problem, encompassing phenomena ranging from multiple coexisting forms of meaning equivalence [2] to untranslatability [36]. For example, the Arabic term بِنْت الخَال, meaning "*daughter of the mother's brother*", is equivalent to the Persian دختردائی, whereas the Chinese term 堂妹, meaning "*younger female patrilineal cousin*", has no equivalent in English.

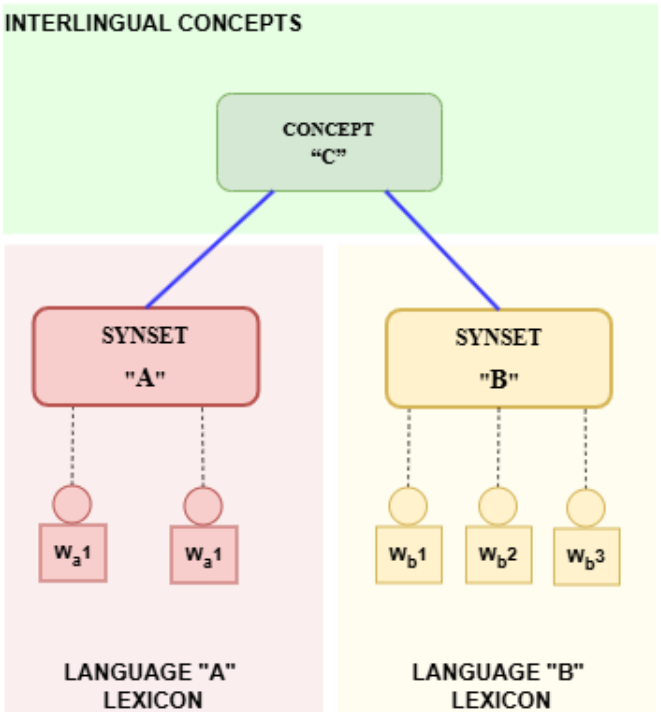


Figure 2.2: Mapping of word meanings in a multilingual lexical resource

Traditional bilingual dictionaries are practical tools designed for the gen-

eral public but often lack fine-grained theoretical precision in mapping meanings across languages [69]. They frequently suggest quasi-equivalences or broader target meanings when exact matches (word senses) are unavailable and may contain lexical gaps—instances where the target language does not lexicalize the meaning of the source word—using free-text definitions. These dictionaries are inherently asymmetric, as they feature distinct source and target languages, reflecting the fundamentally non-reciprocal and context-dependent nature of translation [2].

In the context of multilingual databases, the principle of asymmetry is rarely observed in practice due to scalability concerns. If a multilingual database supports n languages, mappings would require to be defined for $n(n-1)$ language pairs, which is computationally impractical. To address this, all multilingual databases employ a hub (or pivot) meaning representation, known as a *concept*, to which all lexicons of the supported languages are mapped (see Figure 2.2). Concepts are language-independent units that serve as bridges across languages; to be part of the meaning representation layer, each concept must be lexicalized in at least one language. Giunchiglia et al. [67] formalized the equivalence relation between two synsets (S_a and S_b) in a multilingual database as a symmetric and transitive relation. In this formulation, S_a represents a synset from the lexicon of language A, while S_b represents a synset from the lexicon of language B, as illustrated in Figure 2.2. The relation is expressed as follows:

$$S_a \leftrightarrow c \leftrightarrow S_b \Rightarrow S_a \leftrightarrow S_b \tag{2.1}$$

EuroWordNet [154] is one of the most significant and widely recognized early multilingual lexical databases. It extends the structure of the English WordNet to various European languages, creating an interconnected framework that aligns concepts across languages. Similarly, BabelNet [114], which integrates data from PWN and Wikipedia, enables multilingual synset alignments through interlingual links, providing a multilingual resource that supports cross-lingual computational applications.

The Universal Knowledge Core (UKC) [65] is another prominent example of a multilingual lexical database. It is a large-scale resource that, to our knowledge, is the only database that explicitly provides a data model for phenomena such as language-specific words and concepts, as well as cross-lingual lexical gaps. The UKC has been adopted in our work, described in detail in Section 3.4. These resources are indispensable for NLP models to effectively interpret and process text across multiple languages, thereby enhancing accessibility and functionality in a globalized digital landscape.

2.3 The use of lexical-semantic resources

Lexicons play an integral role in numerous NLP applications within artificial intelligence, serving as foundational tools for tasks including natural language understanding, word sense disambiguation, sentiment analysis, machine translation, and cross-lingual information retrieval. A notable example is WordNet, which has garnered over 20,000 citations on Google Scholar[2] since 2020, underscoring its significance and widespread influence across the NLP research community.

The following list presents example efforts from the past five years that utilize lexical resources for various NLP tasks:

1. **Language learning**: The process of acquiring the ability to understand, speak, read, and write in a language. For example, Ono et al. [116] highlights the integration of WordNet to create visual vocabulary maps, showcasing relationships like hypernyms and hyponyms, which aid learners in memorization and comprehension by organizing words based on difficulty, relevance, and frequency. Additionally, Siahaan et al. [137] explores the use of digital lexical tools, including WordNet, Canva, and Kahoot, to improve reading skills in Bahasa Indonesia bagi Penutur Asing (BIPA) programs for non-native Indonesian speakers, demonstrating the potential of technology-driven strategies in language education.

2. **Natural language understanding (NLU)**: A branch of artificial intelligence that enables computers to interpret and make sense of human language, allowing for meaningful interactions between people and machines. For instance, Vij et al. [152] outlines a machine learning method for automating short answer evaluation, where WordNet-based graphs are used to measure semantic similarity between ideal and student answers, resulting in enhanced accuracy and efficiency. Also, Barbouch et al. [17] explores the integration of WordNet embeddings with BERT to improve the model's handling of nuanced lexical semantics. While this approach enhances performance in sentiment analysis and linguistic acceptability tasks, it shows mixed results for sentence similarity and inference tasks, emphasizing the value of WordNet in semantically complex NLU applications.

3. **Word sense disambiguation (WSD)**: The process of determining which meaning of a word is intended in a given context, especially when the word has multiple meanings. For example the following two approaches improved WSD using WordNet. The first study by Karnik et al. [75] enhances homonym detection by integrating WordNet with a modified Lesk algorithm, which improves the accuracy

[2] Accessed on October 30, 2024

of the traditional WSD by leveraging contextual overlaps in glosses and examples to assign context-based weights. The second approach by Loureiro et al. [100] combines neural language models with WordNet to create comprehensive sense embeddings, overcoming issues like meaning conflation in static embeddings. This method maps contextual embeddings to WordNet's semantic structure, yielding significant improvements in identifying word meanings in diverse contexts.

4. **Sentiment analysis:** The process of using NLP to identify and categorize the emotional tone or attitude expressed in a piece of text, determining whether it is positive, negative, or neutral. For example, Ke et al. [78] introduces SentiLARE, a sentiment-aware language model that integrates SentiWordNet data to enhance sentiment analysis by aligning word-level polarity and part-of-speech information with sentence-level sentiment using context-aware attention mechanisms. Tao et al. [143] is another example, which presents a WordNet-guided weakly-supervised approach to extract aspect terms from online reviews, addressing unstructured data and limited annotations by leveraging WordNet's semantic structure to identify context-specific and multi-word terms efficiently.

5. **Machine translation (MT):** The use of computer software to automatically translate text or speech from one language to another without human intervention. Numerous studies highlight the critical role of lexical resources, particularly wordnets, in enhancing MT accuracy and usability. For instance, LexMatcher [164] incorporates bilingual dictionaries to curate high-quality datasets, improving word sense disambiguation and terminology accuracy for MT fine-tuning. Additionally, an unsupervised method [76] utilizes WordNet to simplify Japanese text through refined vocabulary paraphrasing, enhancing accessibility for non-native speakers and achieving high performance on BLEU and SARI metrics. Lastly, DIBIMT [34], a bias-measuring benchmark, reveals challenges in translating polysemous words across multiple languages, underscoring the importance of robust lexical resources in addressing semantic biases and rare word senses in MT.

2.4 LSRs: quality dimensions and key challenges

2.4.1 LSR quality overview

The quality of a LSR is intrinsically linked to the precision and comprehensiveness of its synset components, as well as the robustness of relationships among synset pairs, as demonstrated in our study Khalilia et al. [82]. The structural integrity of these components and relationships is essential to ensuring the reliability of lexical-semantic frameworks, as they underpin

the efficacy of semantic mapping and promote conceptual coherence [81]. High-quality resources facilitate nuanced semantic inferences, enabling NLP models to discern both explicit and subtle relationships within and across linguistic boundaries. Consequently, both monolingual and multilingual lexical resources serve as foundational pillars of the NLP community, driving advancements in context-sensitive language technologies.

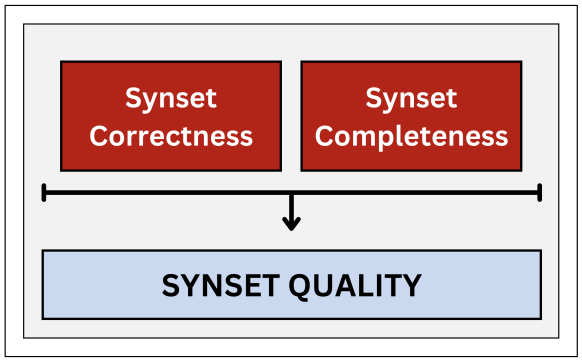


Figure 2.3: The Dimensions of Synset Quality

The quality of a synset is based on two main dimensions, as depicted in Figure 2.3, are synset correctness and synset completeness. The first refers to the correctness of both the information within a synset and its relationships, meaning that synset entries and relations grouped together should accurately represent and describe the same meaning. Accordingly, as demonstrated in our study Khalilia et al. [81], a correct synset should include correct elements [81], as described below:

1. **Correct lemmas**: Synonyms are presented in their canonical form within a set of word forms. For example, "*go*" is the lemma of the words "*go*", "*goes*", and "*went*".

2. **Correct gloss**: A natural language description that conveys the general property (genus) and distinguishing characteristics (differentia) of the concept.

3. **Correct examples**: Includes one or more example sentences that clarify the precise meaning of the described concept.

4. **Correct relations**: Refers to the accurate relations of a synset to other relevant synsets within a LSR, ensuring that relationships (such as synonyms, antonyms, or hierarchical relations) are logically and contextually appropriate.

The second dimension is synset completeness, which refers to the extent to which a synset includes all information necessary to fully represent

a particular concept. A synset is considered complete if it contains both all essential components and relations that convey and describe the same meaning, ensuring comprehensive coverage of the concept it represents. Thus, a complete synset is one with complete lemmas, a complete gloss, and complete examples, as described below. However, a complete synset is defined in our study Khalilia et al. [81], as synset that should include the following:

1. **Complete lemmas**: All expected lemmas of a specific synset should be present in the synset, with no missing synonyms in a given language.
2. **Complete gloss**: A natural language description that includes both parts—genus and differentia—without omissions. The genus corresponds to shared, general knowledge expressed by both the parent and child concepts, while the differentia specifies the unique aspect of the child concept.
3. **Complete examples**: This component contains one or more examples that illustrate the usage of each lemma within the same synset. Examples may be phrases or sentences in a specific language. Synset examples are considered complete if the synset includes at least one example sentence, and, if there is only one example, it uses the preferred lemma (the first lemma in a synset).
4. **Complete relations**: Refers to a synset's comprehensive connection to all relevant synsets within a LSR, encompassing all meaningful relationships, including at least one instance of the expected semantic relations, such as synonyms, antonyms, and hierarchical links.

Given the following English synset example retrieved from the English lexicon in the UKC[3]:

Example 2 *"(n) table, tabular array: a set of data arranged in rows and columns; see table 1."*

This synset is considered a *high-quality synset* because it has accurate and complete components and relationships, as described below:

1. **Synset correctness:** The terms "*table*" and "*tabular array*" accurately and appropriately represent this concept. "*Table*" is the most widely used term, while "*tabular array*" serves as a more technical synonym that emphasizes the organized, grid-like structure. Additionally, the synset gloss "*a set of data arranged in rows and columns*" and the example sentence "*see table 1*" accurately capture the concept of a data table. Furthermore, the relationships of this synset are well-integrated within the broader semantic network, facilitating user navigation among related synsets.

[3]This synset was imported from PWN

2. **Synset completeness:** All relevant lemmas are present in this synset, ensuring no synonyms are missing, such as '*table*", and "*tabular array*". The gloss includes both essential components: the genus "*a set of data*" and the differentia "*data arranged in rows and columns*". Additionally, the synset contains an example sentence that uses the preferred term table. Furthermore, this synset has all expected semantic relationships with relevant synsets across 14 relationships. For example, it is connected to "*data structure*" through a hypernymy relation, to '*multiplication table*" as a hyponym, and to "*list*" or "*chart*" as coordinate terms.

As another example, the following synset is retrieved from the Arabic WordNet (AWN) [51].

Example 3 ضجيج، ضوضاء، ضجج، ضوض, *corresponding to the English synset "noise: sound of any kind, especially unintelligible or dissonant sound; he enjoyed the street noises".*

This synset is considered a *low-quality synset* because it includes some incorrect words and lacks some synonyms. It is deemed incorrect as it incorporates two erroneous lemmas, ضجج and ضوض, which are not found in Arabic dictionaries, such as the Almaany المعاني dictionary[4]. Furthermore, this synset is incomplete, as it lacks the lemma ضجة, meaning "*noise*".

Many synsets in AWN, including the one in Example 3, exhibit significant quality limitations. In our study Freihat et al. [60], we observed that 58% of synsets and 43% of lemmas contained inaccuracies, with 96% of synsets lacking glosses. This issue is pervasive across lexical resources, particularly those translated from PWN, as demonstrated in our survey of LSR quality Khalilia et al. [82].

Accordingly, the quality of a LSR is inherently dependent on the richness and precision of the synsets it contains. The quality of the LSR is determined by the cumulative quality of its individual synsets. LSR quality can be defined as a function encompassing synset correctness and completeness across all its synsets, as described below:

$$Q(R) = \sum_{S \in R} \left(C_{\text{correct}}(S) + C_{\text{complete}}(S) \right) \tag{2.2}$$

Where, $Q(R)$ represents the overall quality of the lexical-semantic resource R, $C_{\text{correct}}(S)$ denotes the cumulative correctness, and $C_{\text{complete}}(S)$ represents the cumulative completeness of R's synsets, S refers to a synset in R, which comprises lemmas L, a gloss G, examples E, and relations R.

$$S = \{\text{L}, \text{G}, \text{E}, \text{R}\} \tag{2.3}$$

[4] http://www.almaany.com/thesaurus.php

The correctness function $C_{\text{correct}}(S)$ is defined as follows:

$$C_{\text{correct}}(S) = f_{\text{correct}}(\text{L}, \text{G}, \text{E}, \text{R}) \tag{2.4}$$

Where $C_{\text{correct}}(S)$ is determined based on the accuracy of each component: L, G, E, and R.

The completeness function $C_{\text{complete}}(S)$ is defined as follows:

$$C_{\text{complete}}(S) = f_{\text{complete}}(\text{L}, \text{G}, \text{E}, \text{R}) \tag{2.5}$$

Where $C_{\text{complete}}(S)$ is determined based on the completeness achieved for each component of the synset.

2.4.2 Key factors affecting LSR quality

As described in the previous section, problems influencing LSR correctness lead to *overload* incorrect entries in this resource, for instance, including excessive of inappropriate senses, incorrect lemmas, or wrong synset relationships, which require additional work, ultimately leading to an overburdened structure. Whereas, issues affecting LSR completeness cause to *underload* correct entries in this resource, such as missing correct senses, accurate lemmas, or appropriate relationships, leaving critical gaps that weaken the resource's overall quality.

Several factors lead to LSR overloading or underloading. These factors can be due to the complexity of language nature, developer (human or machines) mistakes during the annotation process, the gradual change in word meanings over time, or the need to a specialist for the annotation of technical or scientific fields such as medicine, or technology. Some examples of such factors are described below.

1. *Polysemy* is the phenomenon in which a word possesses multiple related meanings, poses significant challenges for LSRs. It introduces semantic ambiguity and can result in errors such as incorrect synset assignments due to the context-dependent nature of word meanings. These challenges are further compounded in compound nouns [58] and hierarchical structures [59], where polysemy complicates the assignment of semantic relations and the management of general-to-specific sense hierarchies.

2. *Incomplete or inconsistent annotation* significantly impacts the quality of LSRs by causing errors in sense assignments, disrupting synset relationships, and omitting key entries (e.g., words or synsets). Variability in annotation quality and inconsistencies across related entries

exacerbate these issues, leading to under-represented information (underload) or redundant and incorrect relations (overload) within the resource [113].

3. *Semantic drift* is the gradual evolution of word meanings over time, presents significant challenges for the development of high-quality LSRs. It can lead to issues such as overloading, where outdated meanings persist, or underloading, where emerging senses are omitted. These challenges are particularly pronounced in fast-evolving domains such as technology and social media[85, 16].

4. *Domain-specific terms* in technical and scientific fields such as medicine, law, or technology, often carry specialized meanings that diverge from their general-language counterparts, posing challenges for their incorporation into general-purpose LSRs. Without appropriate domain adaptation, LSRs risk either omitting these nuanced meanings (underload) or misrepresenting them (overload), especially in specialized disciplines where precise terminology is critical [97].

5. *The granularity of word-sense* [125] is a critical factor in determining the quality of LSRs. Definitions that are too broad can introduce ambiguity by conflating distinct senses, while overly narrow distinctions may complicate the lexicon unnecessarily. Achieving the right balance is challenging but crucial for ensuring both the lexicon's usability and its overall quality.

6. *Linguistic diversity*—the vast range of languages spoken across cultures—plays a fundamental role in shaping not only communication but also the nuances of meaning, cultural expressions, and worldviews embedded within languages. With over seven thousand languages spoken globally, this diversity manifests in countless variations in structure, vocabulary, grammar, and semantics. Each language encodes unique concepts and distinctions that often lack straightforward equivalents in others [65]. For example, in lexical semantics, the English term *cousin*, which signifies "*the child of one's aunt or uncle*", has no exact parallel in Arabic. Conversely, Arabic contains the word عم, meaning a "*paternal uncle*", a concept that lacks a direct translation in English [80]. Such discrepancies highlight how distinct languages capture unique familial or relational nuances, reflecting the social and cultural contexts in which these languages evolved.

   Color terminology also offers further insight into the depth of linguistic diversity. For instance, the Italian word *marrone*, meaning "*chestnut-colored*", lacks a precise equivalent in Persian and Welsh [104]. In Breton, the term *glaz* encompasses a range of hues between blue and green, yet this concept does not directly translate into English or most

other Indo-European languages. These cases illustrate how linguistic diversity gives rise to unique lexical items shaped by local environments, traditions, and perceptual distinctions. Consequently, when languages lack equivalent terms, achieving precise translations or lexical parallels becomes challenging.

Linguistic diversity emerges as a key factor driving cross-linguistic inconsistencies, which arise when aligning resources across languages or developing multilingual LSRs. Structural and semantic differences between languages often hinder direct translation. These inconsistencies complicate the mapping of terms and synsets across languages, frequently resulting in either underload (missing mappings) or overload (incorrect mappings). For example, Piasecki et al. [122] emphasized the importance of ensuring content quality in monolingual resources, particularly those derived from PWN, by incorporating the unique properties of the target language, such as language- and culture-specific concepts. Similarly, Fellbaum et al. [55] observed that direct mappings between a pivot language (e.g., English) and individual languages in a multilingual resource often fail to capture language-specific nuances.

In this thesis, we examine *linguistic diversity* as a key factor influencing the quality of LSRs, whether monolingual or multilingual. Further details on linguistic diversity, including discussions on untranslatability, lexical typology, and its impact on technology—such as cross-lingual NLP applications—are provided in Chapter 3.

2.5 Conclusion

In this chapter, we explored both types of LSRs: monolingual and multilingual. We also examined the need for LSRs in various NLP tasks and natural language learning. Despite the recent attention to deep learning methods and large language models, lexical resources remain integral to NLP applications such as machine translation, word sense disambiguation, and information retrieval.

Resource *correctness* and *completeness* are the two primary dimensions of LSR quality. Both dimensions are sensitive to linguistic diversity, which is a key factor contributing to cross-linguistic inconsistencies. This is particularly evident when aligning PWN with another language or mapping language pairs in a multilingual resource.

3
Linguistic Diversity

Contents

3.1 Introduction

Linguistic diversity encompasses the varied structures, vocabularies, and phonologies found across the world's languages. Understanding this diversity requires insights from *linguistic typology*, a field that classifies languages based on shared structural characteristics rather than genealogical relationships. Linguistic typology provides systematic frameworks for analyzing diversity, enabling researchers to identify universal patterns and language-specific features [42]. Additionally, the language spoken by a community reflects its cultural and social structure. Linguists and lexical typologists have extensively studied these phenomena, particularly in domains such as color terms, geography, body parts, and kinship, with the latter being among the most thoroughly examined [111]. For example, while English has only one term for *cousin*, languages like Arabic and certain South Indian languages have multiple terms—Arabic has eight, and South Indian languages may have as many as sixteen—depending on factors such as gender, age, and whether the relation is on the mother's or father's side. Many local variations across dialects within a single language or across languages within a single country have not yet been fully described or understood. For instance, the term معزوزي *maazoozi* in the Algerian Arabic dialect, meaning "*younger brother*", has no equivalent in Gulf Arabic. Conversely, the Gulf Arabic term ابن العُود *ibn alood*, meaning "*elder brother*", does not exist in

Algerian Arabic, which instead uses the term سِيدي *siedi*.

Beyond linguistic or anthropological interest, the availability of digital resources on language diversity is also essential from a computational perspective. Language processing applications need to be aware of such phenomena of diversity in order to provide high-quality results. For example, a machine translation system needs to tackle cases of lexical untranslatability where a word or expression in a source language has no equivalent in a given target language, and the choice of an approximate translation can change the meaning of an utterance. For instance, for the English sentence *This rice is tasty*, Google Translate provides the Swahili translation *Mchele huu ni kitamu*, which means "*This raw rice is tasty*". This syntactically correct but unintended meaning of eating raw rice occurs due to a *lexical gap*, i.e. a nonexistent equivalent Swahili term for *rice*, as Swahili distinguishes between mchele (uncooked rice) and wali (cooked rice). Such cases of *techno-linguistic bias*—where language technology yields better results for some languages *by design* than for others—tend to remain hidden in monolingual resources but become apparent in multilingual contexts [25, 22].

In recent years, there has been an increasing number of linguistic databases covering a large number of languages. These resources are usually aimed at quantitative studies for comparative linguistics, such as the classification of pain predicates [126], a semantic map of motion verbs [155], the modeling of color terminology [104], the CLICS database of cross-linguistic colexifications [132], DiACL (Diachronic Atlas of Comparative Linguistics), a database for ancient Indo-European languages spoken in Eurasia typology [35], or the Cross-Linguistic Database of Phonetic Transcription Systems [7]. Often, such databases use phonetic representations of lexical units or are limited to a few hundred or a few thousand core concepts, limiting their usability for the processing of contemporary written language. In our experience, most of the existing typology-informed NLP research is restricted to exploring language-specific morphosyntactic features and has ignored diversity within lexical resources [21]. A notable exception is the Universal Knowledge Core (UKC) [65], a large, typology-informed multilingual lexical database that explicitly represents linguistic diversity and is reused in our work.

Our research is part of the *LiveLanguage* initiative, the overarching objective of which is to create, publish, and manage language resources that are "diversity-aware"—i.e. that reflect the viewpoints of multiple speaker communities—and that can be reused by multiple communities: linguists, cognitive scientists, AI engineers, language teachers and students [25]. Contrary to mainstream exploitative practices, LiveLanguage aims to carry out its goals while empowering local speaker communities, giving them control over resources they help to produce [72]. Involving human contributors and deciders from speaker communities is therefore a crucial part of our method.

This thesis examines diversity through three case studies, as outlined in Chapter 1: (1) analyzing language pairs within specific semantic domains, such as diversity in food terms across English-Arabic and Indonesian-Banjarese, and exploring kinship terminology across Arabic-Persian, English-Persian, and English-Turkish; (2) investigating language-independent concepts versus specific languages within a given semantic domain, for instance, comparing kinship concepts with various languages and dialects, including Standard Arabic, Persian, Turkish, seven Arabic dialects, and three Indonesian languages. Additional experiments were conducted on basic-level categories across these languages: Standard Arabic, Turkish, Persian, Indonesian, Javanese, and Banjarese; and (3) focusing on English and a target language, specifically through the translation of PWN synsets into the target language, such as the development of the first diversity-aware AWN.

3.2 Untranslatability and lexical typology

Linguists understand translation from one language to another as a complex and multidimensional problem, ranging from multiple coexisting forms of meaning equivalence to untranslatability [36, 22]. The diversity between cultures is a major cause for this problem appearing on several lexical-semantic levels. Some examples of the linguistic diversity are the richness of Toaripi vocabulary on the various forms of motion verbs describing walking around the beach like *isai* meaning "*go beachward*" and *kavai* meaning "*go inland with respect to the beach*", the language of the coastal Papua New Guinea country, the lack of vocabulary for the word meaning "*sailing*" in Mongolian, which is the language of a landlocked country, or the Arabic word تَسَنَّمَ meaning "*to ascend a camel's hump*".

Lexical typology examines how languages categorize and lexicalize concepts differently, playing a crucial role in understanding untranslatability, where lexical gaps arise due to language-specific conceptual structures. Typological research has identified systematic patterns across various domains, such as kinship, color terminology, and spatial relations. Among these, kinship terms—a primary focus of this thesis—exhibit significant variation across linguistic typologies due to differences in family structures worldwide. For instance, matriarchal societies may employ more nuanced terms for female relatives, whereas strongly patriarchal ones may exhibit a more detailed lexicon for male relatives. In Arabic dialects, kinship terminology not only distinguishes between paternal and maternal brothers but also differentiates blood brothers, full brothers, and breastfeeding brothers. Thus, kinship-related vocabularies vary not only in lexical richness but also in structural organization across languages.

In this research, we focus on lexical untranslatability, which manifests most clearly through the *lexical gap* phenomenon when a word in a source

language does not have a concise and precise translation in a given target language. Lexical gaps are often the linguistic manifestation of culturally or spatially defined specificities of a community of language speakers that cannot entirely be predicted or explained through systematic principles or recurrent patterns [93]. Table 3.1 below presents this phenomenon for nine concepts representing sibling relationships from the kinship domain in eight languages[1]. One can observe that none of the eight languages has concise lexicalizations for all nine concepts, yet each concept is lexicalized in at least one language. Such variations in lexicalization pose a problem for both machine and human translation: for instance, substituting a specific term instead of a broader one may result in injecting unintended meaning. In Javanese, at least four specific terms—(*sedulur*/"*sibling*"), (*adhi*/"*younger sibling*"), (*kangmas*/"*elder brother*"), and (*mbakyu*/"*elder sister*")—are used for expressing the sibling relationship, and accordingly, translating the sentence through Google Translate *my sister is ten years older than me* to Javanese gives the nonsensical sentence *adhiku luwih tuwa sepuluh taun tinimbang aku* meaning "*my younger sibling is ten years older than me*". This result is due to the lack of Javanese vocabulary for the word meaning "*sister*", and also lacks the term meaning "*younger sister*", so the machine translator uses *adhi* meaning "*younger sibling*", which finally produces the semantically absurd output.

Lexical typology studies the diversity across languages according to the structural features of languages with respect to specific semantic fields [123]. Different classical studies are conducted in this field on grammar and phonology, such as VoxClamantis V1.0— a large-scale corpus for phonetic typology [134] and the structure of the space semantic field by identifying a set of semantic parameters and notions depending on the grammatical information of the field's constituents [95]. Other examples of such studies have been conducted on lexical-typological issues that appear across languages during translation, like the presence or absence of lexicalizations in languages. In these articles, authors focused on semantic fields that offer the richness of cross-lingual diversity: family relationships [79], colors [127], food [11], body parts [159], putting and taking events [87], cutting and breaking events [102], or cardinal direction terms [9]. However, only a few open datasets have been published in the scientific research area. These include the classification of kinship by Murdock [111], which has been published in D-PLACE [86]. Part of Kay et al. [77]'s work on colors is published under the lexicon chapter of the World Atlas of Language Structures (WALS) [47]. Additionally, a color categorization dataset by McCarthy et al. [104] is available on GitHub[2].

Digital lexicons have been increasingly used in lexical typology, enabling

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

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

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

Table 3.1: Lexicalizations of nine meanings around the concept of (sibling) in eight languages.

typologists to explore a broader range of languages and semantic domains. One noteworthy example is the KinDiv[3] lexicon [84], which encompasses 1,911 words and identifies 37,370 gaps within the domain of kinship, spanning 699 languages. In our study by Khalilia et al. [80], we extend our investigation into the kinship domain, specifically focusing on linguistic diversity among Arabic dialects and Indonesian languages. Other examples include Viberg [151]'s seminal study on perceptual terminology in 50 languages, later expanded by Georgakopoulos et al. [63] to cover 1,220 languages. Furthermore, the Kinbank database, introduced by Passmore et al. [118], serves as a comprehensive repository of kinship terminology, encompassing more than 1,173 languages and offering a broad coverage of various kinship subdomains. However, there is an important methodological difference between our work [80] and Kinbank's approach to representing terms: Kinbank does not explicitly indicate lexical gaps. For example, our work considers the concept of *son of father's brother as pronounced by a male speaker*, to be a lexical gap in Javanese, while Kinbank maps the Javanese term *sedulur misan*, meaning *cousin*, to this and 95 other meanings. In our approach, *sedulur misan* is recognized as the general term for cousin, with more specific cousin terms identified as lexical gaps. This distinction is useful in comparative linguistics and cross-lingual applications, as the explicit indication of the lack of precise meaning equivalence can be exploited.

Additionally, Concepticon [98] serves as 'a resource for linking concept lists' frequently used in comparative linguistics. The *concept sets* in Concepticon serve the same purpose as the supra-lingual concepts within the UKC resource in our study, which is described in Section 3.4, by providing meaning-based mappings among lists of terms (also known as *concept lists* in Concepticon) across languages.

3.3 The influence of diversity on technology

Linguistic diversity plays a significant role in shaping the performance and accuracy of cross-lingual NLP applications. Languages vary in syntax, semantics, and vocabulary, which poses challenges for NLP systems such as machine translation and word sense disambiguation [72]. Failure to account for linguistic diversity often results in errors and inaccuracies, limiting the effectiveness of these technologies across languages with distinct linguistic structures.

A key issue in cross-lingual NLP applications is lexical untranslatability, which can lead to altered intended meanings. For example, Google Translate has mistranslated the phrase "*do not give cider to your child*" into Arabic as لا تعطي عصير التفاح لطفلك, meaning "*do not give apple juice to your child*". This error arises from the absence of a specific word for "*cider*" in

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

Arabic, highlighting a lexical gap. Similarly, the Arabic word الاحمران, which means "*bread and meat*", has been mistranslated into English as "*red wine*", demonstrating another instance of untranslatability due to the absence of an equivalent term in English [83].

Machine translation systems also struggle with sentences that contain culturally or linguistically specific words. For example, translating "*My brother is three years younger than me*" into Hungarian, Korean, Japanese, or Mongolian often results in syntactically correct but semantically incorrect translations. This occurs because these languages may lack a direct equivalent for "*brother*" or use terms that are uncommon. In Hungarian, for instance, the word *fiútestvér*, which means "*brother*", is relatively rare. Machine translation systems often rely on word frequencies in their training data, leading to errors, such as selecting the term *bátyám*, which means "*my elder brother*", resulting in the nonsensical translation, "*My elder brother is three years younger than me*". This highlights how cross-linguistic differences can lead to translation errors [84].

These challenges are evident in ChatGPT's translation errors. For example, like Google Translate, ChatGPT-4 mistranslated the phrase "*do not give cider to your child*" into Arabic as لا تعطي عصير التفاح لطفلك, producing the same incorrect result. Another example of ChatGPT's translation errors involves regional or culturally specific words. For instance, the Indonesian word *kembili*, which refers to a root vegetable similar to a potato, is mistranslated by ChatGPT-4 as "*umbi-umbian*", a term that refers to a broader class of tuberous vegetables. This error occurs because Banjarese, the regional language, lacks a more specific word for *kembili*. Similarly, in the reverse direction, ChatGPT-4 mistranslates the Banjarese word *palajau*, which refers to a local fruit, as *ketiau*, a word referring to a completely different plant species. These mistranslations stem from lexical gaps, where the target language does not have a precise equivalent for culturally or regionally specific terms [83].

Word sense disambiguation also faces challenges in languages with different semantic distinctions. A well-known example involves the English word *rice*. For instance, the English sentence *This rice is tasty* was machine-translated into Swahili as *Mchele huu ni kitamu*, meaning "*This raw rice is tasty*". This mistranslation occurs because Swahili distinguishes between cooked rice (*wali*) and raw rice (*mchele*), a nuance not present in English [23].

These translation errors are not isolated cases but reflect systematic challenges that arise in languages with high linguistic diversity. Cross-lingual NLP applications, such as machine translation, frequently struggle with linguistic gaps, leading to semantic errors. Similarly, word sense disambiguation systems often fail to distinguish between context-specific meanings across languages, resulting in inaccuracies [24].

In conclusion, linguistic diversity presents challenges for cross-lingual

NLP applications, including lexical untranslatability, context-specific word sense disambiguation, and syntactic-semantic mismatches. Addressing these issues in NLP model design is crucial for enhancing the accuracy and adaptability of technologies such as machine translation across a broader range of languages. Typological lexicons and databases, such as the UKC [65], WALS [47], and Glottolog[4], serve as valuable resources for training and evaluating NLP models that account for lexical diversity. For instance, in this thesis—particularly in Section 11.5—the UKC has been integrated to enhance both machine translation performance and evaluation.

3.4 Universal Knowledge Core

This section describes the Universal Knowledge Core (UKC)[5], a large typology-informed multilingual lexical database that we adopt for the production of diversity-aware datasets in this research [65]. The use of the UKC is motivated by its ability to represent linguistic unity and diversity explicitly: conceptualisations shared across languages, word senses appearing only in certain languages, shared lexicalisations (e.g. cognates), as well as lexical gaps. The theoretical underpinnings of the lexical model of the UKC have been described in Giunchiglia et al. [66] and in Bella et al. [23], and are illustrated in Figure 3.1.

The UKC is divided into a *supra-lingual concept layer* (as shown at the top of Figure 3.1) and the layer of individual lexicons (at the bottom of Figure 3.1). The concept layer includes hierarchies of concepts that represent lexical meaning shared across languages. *Concepts* are language-independent units and act as bridges across languages, and each one should be lexicalized by at least one language to be present in the concept layer. Supra-lingual concepts and their relations (e.g. hypernymy, meronymy) are in part derived from third-party resources such as Princeton WordNet (PWN) [106], and are in part proper to the UKC. In particular, the UKC contains a formal conceptualization of various semantic domains. For instance, it includes an extensive formal conceptualization of kinship domain terms, computed from the KinDiv database, which spans approximately 200 distinct concepts[6]. KinDiv itself, in turn, is based on ethnographic evidence from 699 languages [84]. The UKC is the most comprehensive lexicon available to our knowledge, which motivated our choice of the UKC as a platform for our research.

The *lexicon layer* consists of language-specific lexicons that provide lexicalizations for the concepts from the supra-lingual concept layer, while also asserting *lexical gaps* whenever lexicalizations are known not to exist. Lex-

[4] https://glottolog.org/
[5] http://ukc.datascientia.eu
[6] https://github.com/kbatsuren/KinDiv

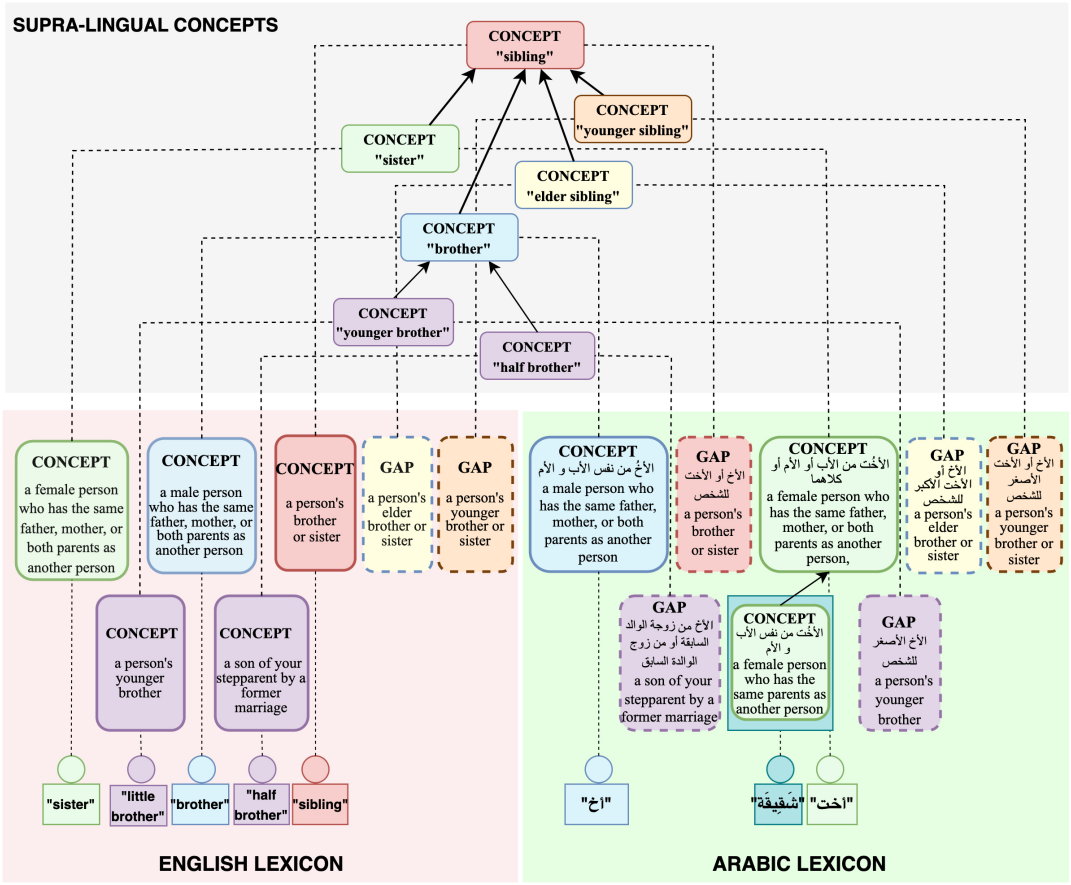


Figure 3.1: Structural elements in the UKC lexical database

icons also provide term definitions as well as lexical relationships specific to the language, such as derivations, metonymy, or antonymy relations. Lexicons can also contain *language-specific concepts* that do not appear in the supra-lingual concept layer. For example, in Figure 3.1, the Arabic شَقِيقَة, meaning "*a female person who has the same parents as another person*", is represented as a language-specific concept. Additionally, by leveraging *linguistic typology*, UKC offers a more nuanced representation of lexical variation. For instance, it differentiates between the English term *sister*, defined as "*a female person who has the same father, mother, or both parents as another person*", and the Arabic-specific term شَقِيقَة, as represented in the respective English and Arabic lexicons. The dual mechanism of defining lexical concepts either on the supra-lingual or on the language-specific level allows for the representation of differing worldviews that would be hard or impossible to reconcile into a single global concept graph. The richness of its lexicon-level linguistic knowledge makes the UKC unique among multilingual lexical databases and particularly suitable for our study.

As mentioned in Section 3.2, a lexical gap for a specific concept is present in a language if there is no concise equivalent word meaning for the concept in that language. For example, neither English nor Arabic has a word meaning *elder sibling*; for such cases, the UKC provides evidence of meaning nonexistence and untranslatability by representing lexical gaps inside lexicons, as shown in Figure 3.1. This information can be used by the NLP community to indicate the absence of equivalent words to downstream cross-lingual applications.

Beyond providing lexical relations between shared word meanings as other multilingual lexical databases do, the UKC also represents a richer set of lexical-semantic connections between language units in a lexicon. For example, the *antonym* lexical relation expresses that two senses are opposite in meaning. While the lexical-semantic relation, *similar-to*, is used to connect two concepts with similar meanings, and the *hypernym-of* connects parent meaning with its child. For instance, in Figure 3.1, the English *little brother* and *brother* are connected through a *hypernym-of* relationship. Such information can be used by the NLP community to indicate the concise equivalent language-specific word meaning to downstream cross-lingual applications, e.g., as the position of a language-specific meaning in a language hierarchy in a lexicon.

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

The major difference between the UKC and other databases, namely Concepticon [98] mentioned in Section 3.2, is as follows: as of mid-2023, Concepticon consists of nearly 4,000 concept sets, principally targeting core vocabularies (basic-level categories) that are the main subject of study in historical and comparative linguistics. Concepticon is under continuous development and has more recently evolved from a flat list of meanings to a hierarchy with broader–narrower relations. At the time of writing, the semantic domains that we are focusing on in this thesis are partially represented in Concepticon, such as the kinship domain: while sibling or grandparent relations are widely covered, fine-grained cousin relationships are mostly missing from it. The UKC, which contains over 100,000 supra-lingual concepts and a wide range of lexical and lexico-semantic relations, is a more suitable resource for our study due to its more complete coverage of the kinship domain and its explicit support for representing term untranslatability via lexical gaps.

3.5 State of the art: development of diversity-aware lexical databases

Multilingual computational applications being in the core of our focus, we also review relevant resources from computational linguistics. For NLP applications, the most popular and widely-known representation of lexico-semantic knowledge is that of *wordnets* that follow the general structure of the original English *Princeton WordNet* [106].

The *wordnet expansion* approach by Fellbaum et al. [55]—an expert-driven lexicon translation effort—is frequently used to produce new wordnets for lower-resourced languages: this approach consists of 'translating' (i.e. finding lexicalizations for) English WordNet concepts ('synsets' in wordnet terminology) into the target language. While this is a straightforward approach that produces resources that remain cross-lingually linked, its downside is that the *translation approach* cannot involve concepts and words specific to the target language and not present in the source language (which in most cases is English). In cases of diverse conceptualisations of the world, the translation approach often results in incorrect approximations. To take the example of Arabic, both versions of the Arabic Wordnet [51, 1] map the English synset of *uncle* ("*the brother of your father or mother; the husband of your aunt*") to the Arabic synset of عَمّ, which means "*the brother of your father*."

A similar situation is observed for Indonesian. As far as we know, the only Indonesian Wordnet currently accessible is Bahasa Wordnet—a bilingual Wordnet for standard Indonesian and Malay languages [115]. It was formed by merging three different wordnets (one in Indonesian and two in Malay) developed mainly by the same expansion approach from PWN. Due to this approach, many English words that have no equivalents in Indonesian are incorrectly mapped, resulting in meaning loss. For example, in Bahasa Wordnet, the English word *sister*, which means "*a female person who has the same father, mother, or both parents as another person*," was mapped to the Indonesian word *kakak* which means "*elder sibling*."

MultiWordNet was an early effort at improving the representation of linguistic diversity in multilingual lexical databases [120]. It is a multilingual lexicon that was built using the *merge* method that, contrary to the translation-based expand approach presented above, maps together existing high-quality bilingual dictionaries. MultiWordNet explicitly represents lexical gaps in its Italian and Hebrew wordnets: about 1,000 in Italian and about 300 in Hebrew [28, 117]. MultiWordNet, however, is a discontinued effort that does not cover the semantic domains that offer richness in linguistic diversity, like kinship, food, colors, and body parts, where it adopts general domains and thus was not suitable for our purposes.

Despite the significant costs and time constraints, the expert-sourced ap-

proach remains the most established and widely utilized method in lexical semantics research for constructing culturally aware lexical resources. For instance, Bella et al. [26] employed an expert-sourced expansion methodology in developing the Scottish Gaelic WordNet. A total of 10,583 English synsets from the Princeton WordNet were translated and validated by Gaelic language experts and subsequently merged with entries from the existing Wiktionary-based Gaelic WordNet. This updated version of the Scottish Gaelic WordNet now comprises over 15,000 synsets and identifies more than 600 Gaelic lexical gaps, representing English words without direct Gaelic equivalents.

Chandran Nair et al. [38] proposed a three-step semi-automatic approach, focusing on WordNet Domains [101], to create the IndoUKC—a multilingual lexical database encompassing 18 Indian languages. Their work aimed to formally capture the words and meanings unique to Indian languages and cultures. By reusing content from the IndoWordNet [44] and introducing a new model for cross-lingual mapping of lexical meanings, the IndoUKC facilitated a diversity-aware representation of lexical meanings. Beyond extensive syntactic and semantic cleaning, the resulting IndoUKC database comprises 194,237 synsets and identifies 119,565 lexical gaps across the 18 Indian languages.

While the aforementioned efforts have made significant strides in exploring and representing linguistic diversity, these methods exhibit several limitations: *(i)* they remain biased toward high-resource languages with well-documented typologies, particularly English, as they often use it as the source language for translation (e.g., from Princeton WordNet to a target language), which can introduce an English bias [72]; *(ii)* expert-driven methods typically focus narrowly on exploring lexical gaps in a limited subset of advanced-level (high-resource) languages, leaving a significant portion of the world's 7,000 languages [66] unexplored; and *(iii)* low-resource languages are especially underserved due to the lack of comprehensive lexical databases and the limitations of automated methods in capturing their linguistic nuances. This thesis seeks to address these biases by introducing a systematic methodology for exploring linguistic diversity without relying on high-resource languages (e.g., avoiding an English-centric approach) and by expanding *typologically* diverse lexical datasets for underrepresented languages.

3.6 Conclusion

This chapter provides an overview of untranslatability and lexical typology, highlighting how integrating linguistic typology into the study of linguistic diversity strengthens our understanding of cross-linguistic variation. Additionally, it explores lexical diversity and introduces the UKC, a lexicon that

formally addresses this diversity across languages by representing language- or dialect-specific concepts and identifying linguistic gaps. Furthermore, it examines the impact of linguistic diversity on technology, particularly in cross-linguistic NLP applications such as machine translation in diversity-rich domains. Typological insights enhance these applications by addressing lexical gaps and word variations. Resources such as UKC [65], WALS [47], and Concepticon [98] play a crucial role in developing diversity-aware lexical databases, advancing both theoretical linguistics and computational applications. Finally, the chapter presents state-of-the-art approaches for developing datasets that account for linguistic diversity.

Part II

Expert-Sourcing Lexical Diversity

4

Expert-Driven Methodology

Contents

This chapter presents a general method employed to collect and produce lexicalizations and identify gaps, with the collaboration of native speakers and language experts. The method described below was applied independently to each of the following case studies, as detailed in Chapter 5: (1) lexicalization of kinship concepts in seven Arabic dialects and three Indonesian languages; (2) The enrichment of the KinDiv resource with kinship terms gathered from ten languages; (3) lexicalization of basic-level categories in Standard Arabic, Turkish, Persian, Indonesian, Javanese, and Banjaraese; and (4) enhancing the Arabic WordNet by improving its content quality. This chapter serves as a tested framework for gathering lexical data in a diversity-aware manner, which we plan to adapt and reuse in future lexicon development projects.

We leverage the UKC to incorporate language-independent concepts (e.g., kinship concepts) as an input dataset for our method and use its data representation model to formalize our data. For instance, in our kinship case study, we utilize an extensive and well-formalised hierarchy of 184 kinship concepts from the UKC.

Furthermore, our work expands the UKC by introducing new concepts, lexicalizations, and lexical gaps in languages and dialects that are either absent from or incompletely represented in the UKC. These lexical gaps and items are integrated into the UKC lexicons, while new concepts are added to its (top) concept layer. The process begins with a lexical-semantic expert generating a contribution task. Native speakers then provide contributions—terms with equivalent meanings and gaps—for a specific language or dialect. The contributions are validated through a two-step process: first, language experts review the collected lexical units and gaps for the language,

and second, a lexical-semantic expert evaluates the identified new concepts, not previously included in the UKC. An overview of this method is presented in Figure 4.1.

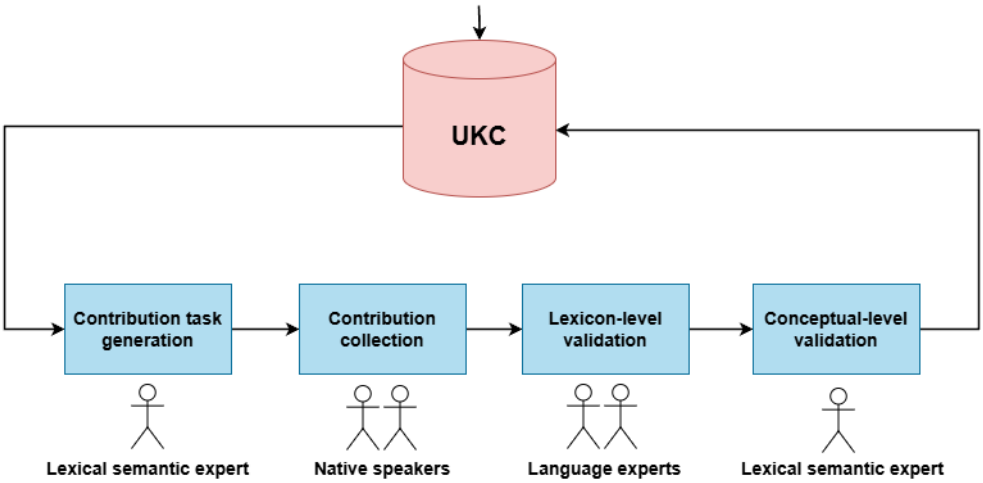


Figure 4.1: Methodology macro-steps and data sources

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

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

4.1 Contribution task generation

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

1. *Dataset of inputs*: Constructing the dataset of general word meanings is the first step of studying diversity across languages and represents the inputs of the contribution collection phase. In this context, the

UKC lexicon is employed to build a dataset, which contains several facilities and tools (e.g., the Exporter tool) that support retrieving categorized data from its interlingual shared meaning layer as introduced in Section 3.4. Moreover, typology datasets or other approaches can be used for that, such as the kinship dataset from Murdock [111] and the color categorization by McCarthy et al. [104]; or gathering data from online dictionaries using automatic methods, i.e. KinDiv [84] retrieves some of its kinship terms from Wiktionary. The constructed dataset is a spreadsheet containing language-independent meanings from one semantic field. At the same time, its content is distributed into subdomains (sheets) for usability and simplicity in designing a concept hierarchy for each subdomain which is a helpful tool for lexical-gap exploration. One spreadsheet row is generated for each concept, containing the concept ID, the source concept definition in the standard language, another definition in English, as well as empty slots for inserting a lexical gap or a word with equivalent meaning, and the data provider's comments in a language or dialect.

2. *Interlingual concept hierarchy*: Modeling the interlingual shared meaning space is essential to explore lexical gaps systematically. In this task, the UKC concept hierarchy is exploited. UKC is the only resource introducing a hierarchy of shared meanings across languages for each semantic field, such as kinship, colors, or food. Furthermore, UKC uses a hybrid linguistic-conceptual approach in modeling each domain. This approach adopts actual domain ontology and linguistic data from typological literature. For example, a fragment of the brotherhood hierarchy in the top layer of the UKC is shown in Figure 3.1. A native speaker can compare each examined concept from the spreadsheet with the hierarchy of its domain to extract additional knowledge about its meaning based on a concept's position in the hierarchy, which helps to provide a concrete answer in terms of a gap or a lexical unit.

4.2 Contribution collection

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

1. They are given the authority to skip concepts, stop contributions, or leave a comment when they deem the terms are becoming too culture-specific and consequently need an exact answer.

2. They are asked to provide a lexicalization in a language (or dialect) that gives meaning equal to the concept's meaning.

3. They are asked explicitly to identify lexical gaps where no language (or dialect) lexicalization exists.

4. Within a language (or dialect) and a subdomain (e.g., cousins), they are asked to provide new concepts that did not exist in the list of inputs which is imported from the UKC by providing a word (lemma) and a clear description of its meaning.

The process of providing such contributions is depicted in two flowcharts; for instance, Figure 4.2 shows the flowchart of the candidate gap (on the left-hand side of the figure) and candidate equivalent word meaning (on the right-hand side of the figure) exploration; it starts identifying a standard language (or dialect) and providing a native speaker with a spreadsheet including a list of subdomain concepts (inputs). Then, a native speaker is asked to find a linguistic resource in the language and use it to search for concepts (concept-by-concept) to confirm lexicalizations and gaps. He/she can use a linguistic resource in the search process as the following steps: searching in a well-known dictionary, then in Wiktionary— a large multilingual online lexicon after that in a typology dataset (if it is available), and finally, using Google search (based on the count of search hits). More details about these steps are described in Section 5.1. The native speaker can rely on search results and the count of Google hits to give a more concrete answer on whether a concept has a lexicalization or is a gap in the examined language; such candidates are passed to the next phase- lexicon-level validation.

A new concept collection is a third contribution in this phase, where the steps of a candidate new concept exploration in a language can be seen in Figure 4.3. A native speaker can examine the list of subdomain concepts and provide his/her (own) concepts with their definitions that he/she believes have not existed in the list. The same search steps in gap identification can be followed in this task. As shown in Figure 4.3, all candidate new concepts are passed to the two subsequent validation phases: lexicon- and conceptual level.

4.3 Lexicon-level validation

Our lexicon-level validation method formally and explicitly addresses individual gap identifications and their quality, as well as equivalent word meanings and new concepts. It allows a qualitative evaluation of the entire list of provided contributions through word-by-word and gap-by-gap in a loop between a native speaker and a validator. A word, a gap, or a new concept does not pass this validation until the native speaker provides the correct answer for each of them, as shown in the flowcharts in Figure 4.2 and Figure 4.3.

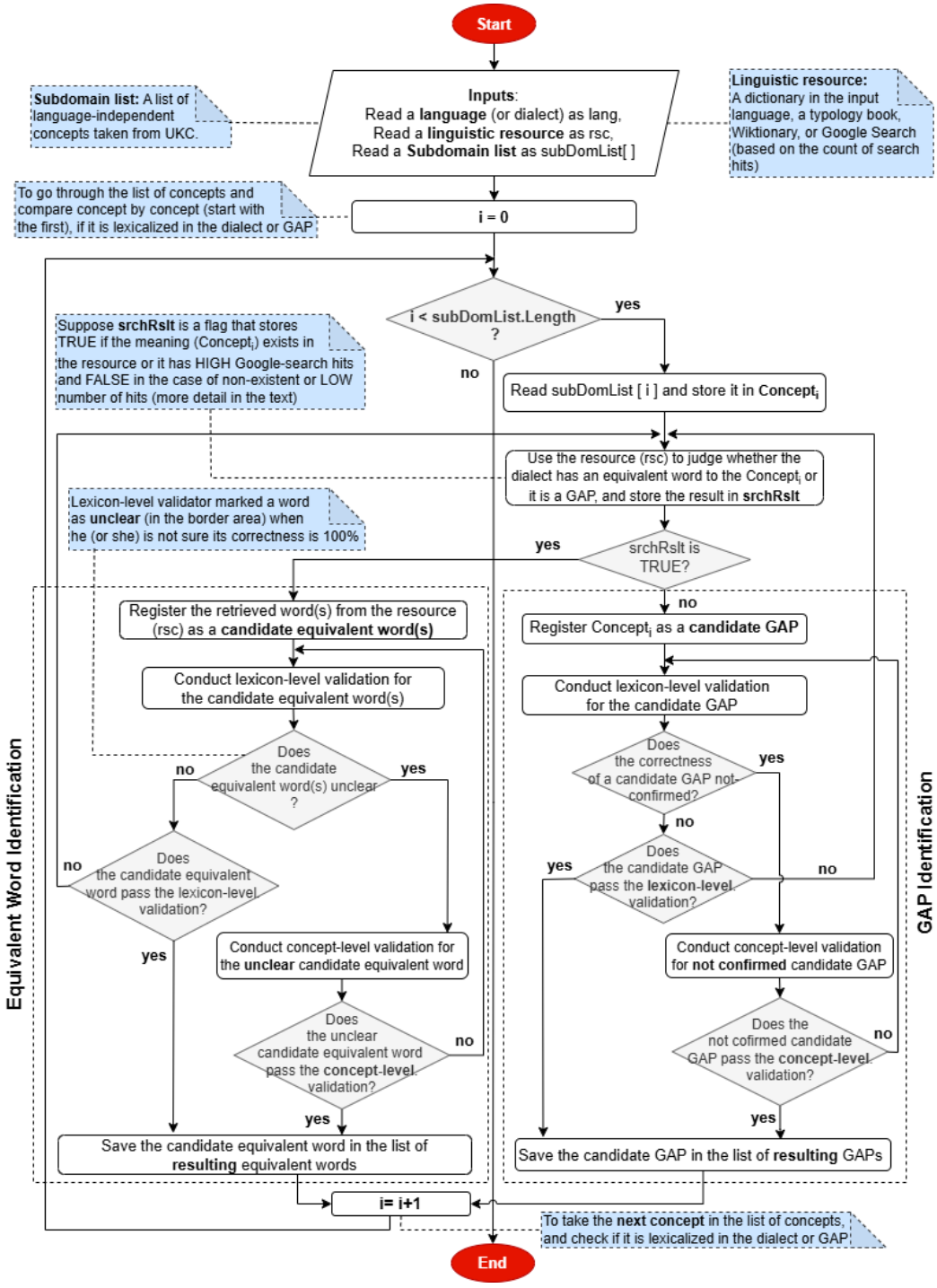


Figure 4.2: Flowchart of gap and equivalent word meaning identification

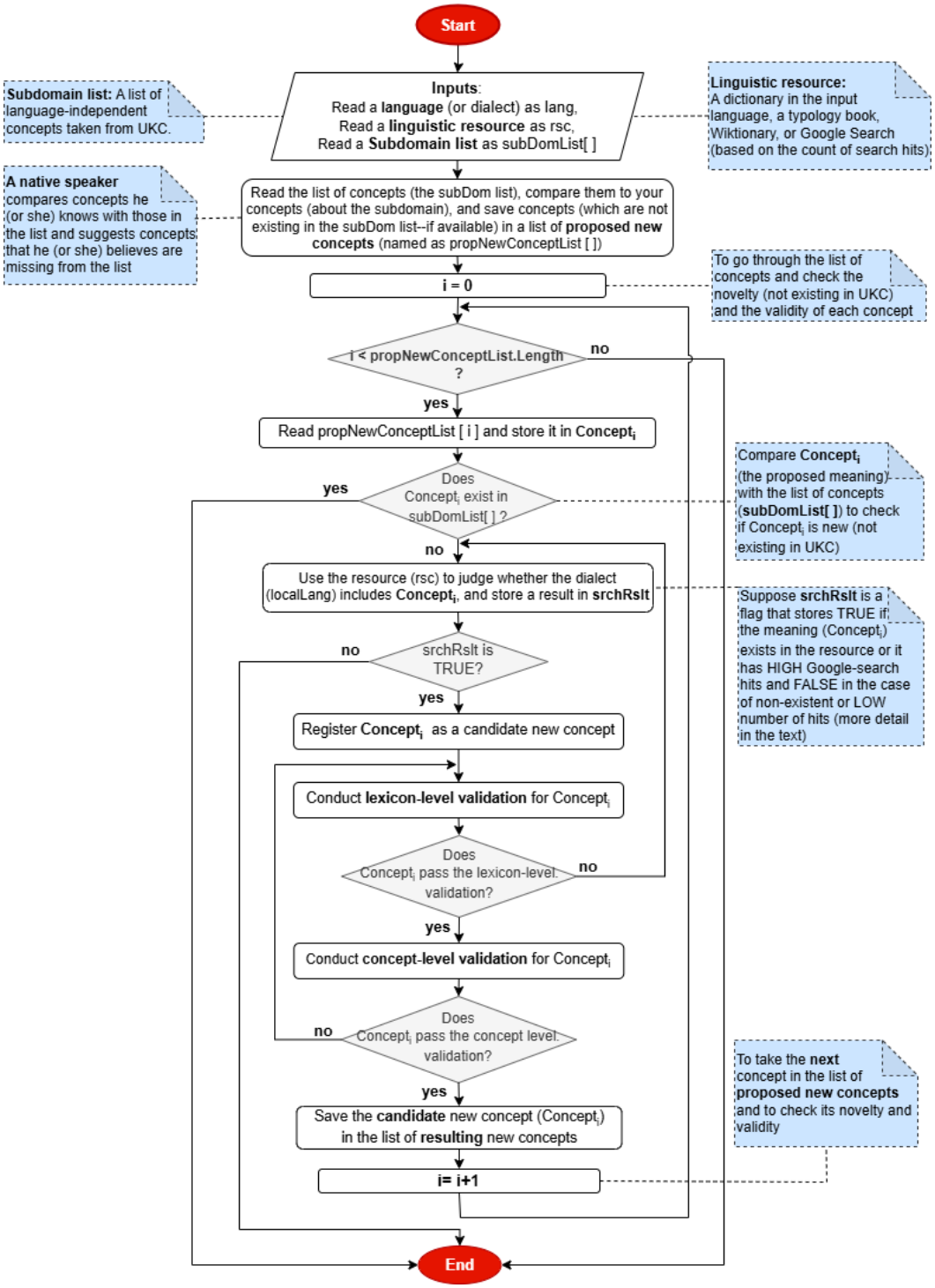


Figure 4.3: Flowchart of a new concept collection

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

1. *Equivalent word meanings*: validate the correctness of all provided words in the language (or dialect) by marking them up as correct, incorrect, or unclear for borderline cases and by providing correct words or indicating them as lexical gaps for incorrect ones.

2. *Lexical gaps*: validate the word meanings marked as lexical gaps by the native speaker in the language, either as confirmed gaps or as non-gaps due to an existing lexicalization in that language, which the validator needs to indicate.

3. *New concepts*: validate all proposed new word meanings in each subdomain by marking them up as correct, correct but not new (in case the supposedly new concepts already existed in the list), or not accepted (in case another concept already existed in the list to express the meaning, or the validator does not consider it as a desirable suggestion for other reasons).

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

4.4 Concept-level validation

In this step, a *lexico-semantic expert*, who manages the UKC system, verifies new concepts and assesses their quality to determine whether they should be accepted or rejected for inclusion in the supra-lingual concept layer. Additionally, the expert resolves ambiguities related to unclear words and borderline cases of unconfirmed gaps or non-gaps. A *lexico-semantic expert* specializes in analyzing word meanings (semantics) and their structure or

usage within a language (lexicon). Their work involves examining relationships between words, interpreting meanings, and understanding their roles in communication.

This validation is based on a discussion session with the language expert responsible for lexicon-level validation through concept-by-concept and case-by-case issue validation. A spreadsheet containing all new concepts and determined (words and gaps) to be examined is used. Columns of this sheet are the same columns in the previous step and two additional empty ones: the evaluation and concept-level validator's comment. The following tasks are used:

1. *New concepts*: Validate all proposed new concepts in each subdomain by marking them up as correct, correct but not new (in case the supposedly new concepts already existed in the UKC), or not accepted (in case another concept already existed in the UKC to express the meaning, or the validator does not consider the new concept as a desirable suggestion for any other reason).

2. *Unclear words*: Validate the correctness of unclear word cases considered in the border-area by the lexicon-level validator by marking them as correct or incorrect and writing a comment.

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

Correct new concepts are imported into UKC by merging them with the supra-lingual conceptual layer. In contrast, not-accepted ones and those correct but not new are returned to the validator at the lexicon level, who may also return them with a comment describing the reason to the native speaker to address an included problem. In a new cycle, modified new concepts by the native speaker are transferred to this phase through the validator of lexicon-level; then, the validator at this level reviews the updates and decides whether to finish the revision cycle by accepting or rejecting the new concepts or issue a new one for more review, as shown in Figure 4.3. In addition, confirmed words and gaps output from this step are integrated with the language lexicon in the UKC, as shown in Figure 4.2.

5 Expert-Sourcing Experiments

Contents

This chapter demonstrates the validation of the expert-driven methodology introduced in Chapter 4. It involves a series of experiments designed to explore and enhance understanding of diversity and semantic representation in different semantic domains (e.g., kinship terms and basic-level categories) across various languages, and dialects. The first two sections examine the richness and variability of kinship terms in Arabic dialects and Indonesian languages, highlighting their linguistic nuances. Section 5.1 focuses on diversity among Arabic dialects, while Section 5.2 addresses it across Indonesian languages. Building on these insights, Section 5.3 presents the expansion of KinDiv, a multilingual lexicon dedicated to kinship, which integrates diverse linguistic data from 699 languages. In Section 5.4, an experimental study explores the representations of basic-level categories across six languages—Arabic, Persian, Turkish, Indonesian, Banjarese and Javanese—emphasizing cross-linguistic differences. Finally, Section 5.5 presents an experiment aimed at improving the Arabic WordNet, focusing on enhancing its content quality to better represent Arabic culture.

5.1 Lexical diversity in kinship across Arabic dialects

This section demonstrates the use of the expert-driven methodology on kinship terminology from seven Arabic dialects. Arabic is the official language of more than four hundred million native speakers in twenty-two countries in the Middle East and northern Africa. Classical Arabic or Modern Standard

Arabic (MSA) refers to the standard form of the language used in academic writing, formal communication, classical poetry, and religious sermons [51]. Surprisingly lexical diversity is manifested between Arabic dialects, evident in our study between seven of the twenty dialects spoken worldwide. The selected dialects are Egyptian, Moroccan, Tunisian, Algerian, Gulf, and South Levantine (two examples: Palestinian and Syrian). Let us take the example of the Gulf word الخَال العودْ meaning "*mother's elder brother*," which has no equivalent in South Levantine or Moroccan; instead, they use the more general word الخَال meaning "*mother's brother*," which can be used for both meanings "*mother's younger brother*" or "*mother's elder brother*". In this thesis, we perform an experiment on the Arabic dialects to capture their diversity in the kinship domain. The resulting dataset with dialect-specific kinship terms will be integrated with an instance of the Universal Knowledge Core for Arabic (Arabic UKC)[1] ongoing project, which is the first diversity-aware lexical resource for Arabic dialects so far.

As mentioned in Section 3.4, the UKC resource is our data source in building the input dataset of kinship-independent language concepts and formalizing such concepts and new word meanings (not existing in the inputs) explored in this experiment. For example, the brotherhood hierarchy is shown in the top layer of the UKC in Figure 3.1. In this study, contributions are provided by seven native speakers (one per Arabic dialect). Regarding the contributors' socio-linguistic background, each has at least a master's degree and was born and educated, at least up to high school level, within the native speaker community. The participants' linguistic backgrounds are presented below:

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

[1] http://arabic.ukc.datascientia.eu/concept

Seven experiments (one for each dialect) are performed to explore lexical units and gaps using the expert-driven method. In each experiment, a spreadsheet of kinship concepts is imported from the UKC (as the source, they were computed from the KinDiv database, described in Section 5.3), which serves as an input dataset to the contribution (diversity-aspects) collection step. These kinship domain concepts are language-independent units representing lexical meaning shared across 699 languages and spanning 184 distinct concepts. UKC categorizes kinship concepts into six groups; each one contains a distinct subset of concepts sharing a common kinship type meaning called a subdomain, for example, sibling and cousin subdomains. The spreadsheet (the dataset) consists of six sheets, and each one represents a kinship subdomain. See Table 5.1, which shows the subdomain names and the count of containing concepts per subdomain of the dataset.

Subdomains	**Count of Concepts**
Grandparents	19
Grandchildren	27
Siblings	21
Uncle/Aunt	27
Nephew/Niece	33
Cousins	57
Total	184

Table 5.1: Statistics on the concepts in the input dataset.

In the contribution collection, a native speaker answers by filling a lexical unit or gap in a row empty slot specified for each concept. Linguistic resources and Google Search are used to provide answers as precise as possible. For example, the المعاني Almaany dictionary, Wiktionary[2], and the *Fiqh AlArabiyya* typology book [112] are employed in sequential steps to give a judgment on cousin words in Syrian. Additionally, counting the number of hits returned by the Google search engine is another helpful indicator, where a high count of hits indicates a searching word (i.e., ابن العمة meaning "*son of father's sister*," has 131.5 million hits) is a lexical unit in Syrian. In contrast, a low count indicates a lexical gap; for example, الخؤولة meaning "*maternal cousin*," has 158 thousand hits. Google hits of other cousin terms are shown in Table 5.2. Since Arabic words can be written and read with or without diacritics (i.e., *fatha* above a letter or *kassra* under it), thus, each word is typed in two forms. Note that the content of this matrix cannot be considered the only criterion for gap exploration because word hits may contain a count of other hits resulting from searching in other Arabic dialects for the same word.

[2] http://ar.wiktionary.org

Concept	With/Without Diacritics	Hit Counts	
العمومة Paternal cousin	العُمومَةُ العمومة	1.94M 1.1M	3.04M
الخؤولة Maternal cousin	الخُؤُولَةُ الخؤولة	111K 47K	158K
ابن العم Son of father's brother	اِبْن العَم ابن العم	84.8M 9.16M	93.96M
بِنت العم Daughter of father's brother	بِنْت العَم بِنت العم	8.43 M 74.7 M	83.13 M
ابن العمة Son of father's sister	اِبْن العَمَّة ابن العمة	12.5 M 119 M	131.5 M
بِنت العمة Daughter of father's sister	بِنْت العَمَّة بِنت العمة	9 M 21.4 M	30.4 M
ابن الخال Son of mother's brother	اِبْن الخَال ابن الخال	5.61 M 27.4 M	33.01 M
بِنت الخال Daughter of mother's brother	بِنْت الخَال بِنت الخال	3.99 M 26.7 M	30.69 M
ابن الخالة Son of mother's sister	اِبْن الخَالَة ابن الخالة	12.5 M 4.09 M	16.59 M
بِنت الخالة Daughter of mother's sister	بِنْت الخَالَة بِنت الخالة	11 M 5.67 M	16.67 M

Table 5.2: Count of Google search hits for cousin concepts in Arabic

The overall contribution collection effort resulted in the identification, formalization, and collection of 180 words, 1,108 lexical gaps, and 19 new concepts in this experiment.

Validation was carried out in two phases. In the first phase, words and gaps were validated at the lexicon level by the author of this thesis, a Ph.D. student specializing in lexical semantics and a native speaker of Arabic, along with a university instructor who is also a native Arabic speaker with expertise in linguistic semantics and a good knowledge of Arabic dialects. In the second phase, new concepts were verified and approved for inclusion in the concept layer of the UKC by another university instructor,

a lexical-semantic expert, and the UKC system manager.

Using the lexicon-level validation method, the author evaluated the collected data in Palestinian and Syrian, while the university instructor validated the remaining five dialects. Results can be seen in Table 5.3, whereby correctness, we understand the number of words (or gaps) validated as correct divided by the total number of words (or gaps). In the case of an incorrect word, the validator either provides a correct word or indicates it as a lexical gap. For example, for the Algerian dialect, the correctness of gathered words is 85.71% and that of gaps is 98.08%. Four Algerian words were deemed incorrect: مانّي for the meaning "*maternal grandmother*", لالّة for the meaning "*paternal grandmother*", جَدّ for the meaning "*grandfather*", and باب الشيخ for the meaning "*grandparent*". The validator indicated "*maternal grandmother*", "*paternal grandmother*", and "*grandparent*" as gaps, while he replaced the mistaken word جَدّ with the correct word باب الشيخ for "*grandfather*". For gap evaluation, the linguistic expert validates a lexical gap by confirming it as a gap or as a non-gap due to an existing word in a dialect, for which he must provide the correct word. For instance, *Participant 1* identified the meanings "*elder sister*", "*father's elder sister*" and "*mother's elder sister*" as gaps in Algerian, but the validator did not accept them and provided the polysemous word لالّة for each of them. Evidence for validation was obtained from the dictionary *Dictionnaire arabe algérien*[3] and from usage attested in Algerian TV films. Upon discussion between the validator and the participants, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts provided in MSA and English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

Dialects	**Correctness of Native Speaker Contribution**	
	Words	**Gaps**
Algerian	85.71%	98.08%
Egyptian	96.90%	97.37%
Moroccan	95.83%	97.53%
Palestinian	100%	98.76%
Syrian	91.67%	95.00%
Tunisian	95.65%	98.14%
Gulf	100%	96.79%
Average	**95.11%**	**97.38%**

Table 5.3: Validator evaluation of words and lexical gaps by dialect

[3] https://www.lexilogos.com/arabe_algerien.htm

5.2 Lexical diversity in kinship across Indonesian languages

This section demonstrates the use of the expert-driven methodology on kinship terminology across three Austronesian languages from Indonesia: Indonesian, Javanese, and Banjarese. Contrary to the Arabic dialects in Section 5.1, these three languages are not mutually intelligible.

Indonesia is the fourth most populous country in the world, and it has more than 700 living languages [49]. The national language spoken in Indonesia is Bahasa Indonesia/Indonesian[4] language, which was decided in the historic moment of Youth's Pledge, October 28th, 1928. However, many Indonesians speak more than one language. For example, out of 198 million people that speak Indonesian, 84 million of them speak Javanese [3].

Even with the high number of speakers, the count of NLP research on Indonesian languages is very low compared to other languages around the world. As of 2020, the count of published papers on the Indonesian language is lower than other languages with less speaker count, such as Polish and Dutch [3]. Not surprisingly, the amount of research on other languages (i.e., Banjarese and Javanese) in Indonesia is much lower than that. It is therefore motivating to conduct this study that discovers the richness of linguistic diversity across three Indonesian languages: standard Indonesian, Banjarese, and Javanese. In one semantic field, kinship, we have found that diversity is manifested in these languages; for example, in Javanese, the word *ponakan jaler* meaning "*nephew*", is a lexical gap in Banjarese, and in the opposite direction, the Banjarese *gulu* meaning "*parent's second eldest sibling*" is also a gap in Javanese.

As in the Arabic experiment, we use the UKC lexicon to create the input dataset of kinship terms, which are independent language and formalizing such terms and also new concepts (not existing in the input dataset) identified in this experiment, as shown in the top layer of the UKC in Figure 3.1 for the brotherhood categorization.

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

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

[4]Throughout this thesis, the term "Indonesian" is used to refer specifically to "Standard Indonesian".

For each language, an experiment was carried out to identify words and gaps associated with the same 184 kinship concepts as in the Arabic study (see Table 5.1). For example, in Banjarese, the dictionary *Kamus Bahasa Banjar Dialek Hulu-Indonesia* [15] and Google Search hits were used in subsequent steps to provide a precise answer on each concept from the given list of inputs. Such search steps were also followed by the Banjarese native speaker for the task of judgment on new concepts identified in the uncle/aunt subdomain. For instance, the Banjarese term *gulu*, expressing an uncle/aunt relationship with the meaning of "*a parent's second eldest sibling*" and attested by the dictionary above, did not previously exist in the UKC or in the KinDiv dataset (described in Section 5.3), nor in Murdock [111]. Indonesian and Javanese native speakers also follow the same steps and use the dictionaries of Badan Pengembangan dan Pembinaan Bahasa [13] and Utomo [149] for the task of judgment on terms and gaps identified in Indonesian and Javanese, respectively.

The overall contribution effort yielded 41 words and 517 lexical gaps in Indonesian, Javanese, and Banjarese. Additionally, three previously unattested word meanings were identified and formalized as new concepts in this experiment.

For validation, a two-step process was conducted in this study. The first step involved validating words and gaps contributed by native speakers, performed by a master's student who is a native Indonesian speaker with a good command of all three languages. The second step focused on concept-level validation, carried out by a university instructor, a lexical-semantic expert, and the UKC system manager. In this step, new concepts were reviewed and approved for inclusion in the concept layer of the UKC.

Table 5.4 provides correctness results over native speaker contributions, provided by the validator. Upon discussion between the validator and the contributors, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts, provided in English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

Languages	**Correctness of Native Speaker Contribution**	
	Words	**Gaps**
Indonesian	90.91%	98.27%
Javanese	94.44%	95.78%
Banjarese	91.7%	97.67%
Average	**92.35%**	**97.24%**

Table 5.4: Validator evaluation of words and lexical gaps by language

5.3 Enrichment of KinDiv: a multilingual lexicon for kinship

Our efforts in this experiment were part of the construction of the KinDiv lexicon project [84]. This lexicon encompasses 1,911 words and 37,370 gaps within the domain of kinship, spanning 699 languages. A hybrid approach was employed in the construction process, relying on three sources of evidence for lexicalization: Murdock's lexicalization patterns [111], Wiktionary entries, and direct input from native speakers. The latter source represents our primary contribution in this research. Accordingly, these sources are described as follows:

1. Lexicalization patterns were provided for 566 languages, retrieved from Kemp et al. [79]. To give an example of a lexicalization pattern, the Javanese language follows the *Algonkian sibling pattern*, meaning it lexicalizes the sibling terms *elder brother*, *elder sister*, and *younger sibling*. Although this database does not include the actual words, it can be used to infer lexical gaps.

2. From Wiktionary, kinship terms were extracted from Wiktionary for 166 languages, 144 of which were complementary to the Murdock dataset. This was achieved using a custom-written parsing script based on a set of predefined inference rules that automatically identified the presence of gaps.

3. Lastly, for the languages Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish, two native speakers per language (at least one of whom is a university instructor or a PhD student specializing in linguistics) were asked to provide lexicalization and gap information for all kinship concepts. This information was used for automatic-approach evaluation and as a gold-standard source of terms and gaps for languages where incorrect, incomplete or no data was available in Murdock's dataset or Wiktionary.

As mentioned above, in this experiment, we contributed as the third source for the development of the KinDiv lexicon. Specifically, we built a diversity-aware collection of kinship terms across ten languages using the methodology described in Chapter 4. These languages range from high-resourced languages, such as English, to low-resourced ones, such as Kannada. Similar to the Arabic and Indonesian experiments described in Section 5.1 and Section 5.2, we used the UKC to construct the input dataset of kinship concepts (Table 5.1) for the contribution collection step. The UKC was also employed to formalize these concepts, as well as new concepts (not present in the input dataset) identified during this experiment, as shown in the top layer of the UKC in Figure 3.1 for brotherhood categorization.

The overall contribution effort resulted in 396 words and 1,503 lexical gaps provided by native speakers in the ten languages: Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish.

To evaluate the collected data (words and gaps) at the lexicon level, each language was assessed by a language expert with at least proficiency in the language and a good understanding of lexical semantics. This evaluation covered all languages involved in the experiment. For concept-level validation, new concepts were reviewed and approved for inclusion in the UKC's concept layer by a university instructor, a lexical-semantic expert, and the UKC system manager.

Languages	**Correctness of Native Speaker Contribution**	
	Words	**Gaps**
Arabic	100%	100%
English	100%	100%
French	80%	100%
Hindi	100%	100%
Hungarian	100%	99.2%
Italian	100%	100%
Kannada	99%	98.3%
Malayalam	96.7%	99.4%
Mongolian	98.1%	98.6%
Spanish	93.8%	100%
Average	**96.8%**	**99.6%**

Table 5.5: Validator evaluation of words and lexical gaps by language

Subdomains	**Correctness of Native Speaker Contribution**	
	Words	**Gaps**
grandparents	93.3%	99%
grandchildren	98.1%	100%
siblings	91.7%	100%
uncle/aunt	97.9%	100%
nephew/niece	100%	99.4%
cousins	100%	99.4%
Average	**96.8%**	**99.6%**

Table 5.6: Validator evaluation of words and lexical gaps by domain

Validation results by language are presented in Table 5.5, while results

by domain are shown in Table 5.6. Correctness is defined as the number of words (or gaps) validated as correct divided by the total number of words (or gaps). In cases of incorrect words, the validator either provides a correct word or identifies it as a lexical gap. For instance, our French evaluator noted that *adelphe* meaning "sibling", is a newly coined and largely unknown term (e.g., it does not appear in the 1996 edition of the *Robert*). Additionally, *aïeul* and *aïeule* were identified as obsolete for designating "*grandfather*" and "*grandmother*", respectively. Similarly, the Spanish gender-neutral term *hermane* for "*sibling*" was identified by our Spanish evaluator as a rarely used neologism. Consequently, all these concepts were identified as lexical gaps. Furthermore, the French term *cadet* was determined not to mean "*younger sibling*", as suggested by the contributor, but rather "*younger brother*". Apart from these examples, the terms provided by native speakers turned out to be of very high quality, where, the correctness average of each of words and gaps is 100%.

5.4 Experiment: basic-level categories

The concept of basic-level categories, introduced by Rosch et al. [131], refers to the level of categorization that is most natural, efficient, and cognitively salient for humans. These categories represent an intermediate level in a hierarchical structure of concepts, falling between broad superordinate categories (e.g., "furniture") and more specific subordinate categories (e.g., "rocking chair"). For example, when people see a wooden seat with four legs and a backrest, they are more likely to categorize it as a "chair" (basic level) rather than as "furniture" (superordinate) or "recliner" (subordinate). Basic-level categories strike a balance between being general enough to encompass multiple specific examples and specific enough to provide meaningful distinctions.

Rosch [130] indicates that basic-level categories are the first to be acquired by children during language development and are often the default labels people use in everyday conversation. For instance, a child is more likely to learn the word "dog" (basic-level) before learning "animal" (superordinate) or "Golden Retriever" (subordinate). Additionally, basic-level categories are often associated with clear mental images and shared perceptual features across category members, such as the prototypical image of a "bird" being a robin rather than a penguin.

This section demonstrates the application of the expert-driven methodology to basic-level categories across six languages: Arabic, Turkish, Persian, Indonesian, Banjarese, and Javanese. In this study, we utilized the UKC lexicon to construct an input dataset of basic-level categories and formalized 968 concepts included in this dataset. Six native speakers—one per language, all born and educated within their respective language commu-

nities—were recruited as contributors. The participants' linguistic backgrounds are as follows:

1. *Participant 1*: a native Arabic speaker with a good command of English.
2. *Participant 2*: a native Turkish speaker with a good command of English.
3. *Participant 3*: a native Persian speaker with a good command of English and Arabic.
4. *Participant 4*: a native Indonesian speaker with a good command of English and Javanese.
5. *Participant 5*: a native Banjarese speaker with a good command of Indonesian, Javanese, and English.
6. *Participant 6*: a native Javanese speaker with a good command of Indonesian, Banjarese, and English.

For each language, an experiment was conducted to identify words and gaps associated with the same 968 concepts. For example, linguistic resources such as المعاني the Almaany dictionary[5] and Wiktionary were employed sequentially, with Google Search used to ensure precise judgments in Arabic. In Indonesian and Banjarese, resources such as the *Kamus Bahasa Banjar Dialek Hulu-Indonesia* dictionary [15] and Google Search were similarly utilized.

The overall contribution effort resulted in identifying 159 lexical gaps and a total of 5,649 words across Arabic, Turkish, Persian, Indonesian, Banjarese, and Javanese. These findings did not align with our initial expectations, as we had assumed that basic-level categories would not include lexical gaps. However, Persian exhibited 65 gaps. Notably, several common concepts, such as "cow", "bake", and "write", were identified as lexical gaps in Persian. For example, Persian uses گاو to refer to "*both male and female domestic cattle*", and پختن means both to "*cook or bake*". Similarly, in the context of writing, Persian includes context-specific meanings, such as یک نامه بنویس, which specifically means to "*write a letter*".

For validation, this process involved verifying words and gaps contributed by native speakers. The verification was performed by six linguistic experts:

1. *Validator 1*: a native Arabic speaker with a PhD in lexical semantics.
2. *Validator 2*: a native Turkish speaker and master's student with good experience in linguistics.

[5] http://www.almaany.com/thesaurus.php

3. *Validator 3*: a native Persian speaker and PhD student with good experience in linguistics.
4. *Validator 4*: a native Indonesian speaker and master's student with good experience in lexical semantics.
5. *Validator 5*: a university instructor in a linguistics department who is a native Banjarese speaker.
6. *Validator 6*: a university instructor in a linguistics department who is a native Javanese speaker.

Table 5.7 presents the correctness results of native speaker contributions, as evaluated by the validator. Through discussions between the validator and the participants, it was determined that the participants' mistakes often stemmed from misunderstandings of certain concepts, provided in English. The validator ensured these errors were corrected, thereby increasing the overall correctness of the final dataset toward 100%.

Languages	**Correctness of Native Speaker Contribution**	
	Words	**Gaps**
Arabic	98.77%	97.89%
Turkish	99.68%	93.75%
Persian	99.93%	98.44%
Indonesian	97.47%	100.00%
Banjarese	95.18%	95.78%
Javanese	93.82%	90.33%
Average	**97.48%**	**96.03%**

Table 5.7: Validator evaluation of words and lexical gaps by language

5.5 Advancing the Arabic WordNet: elevating content quality.

5.5.1 State of the art: Arabic WordNet

The first effort of building an Arabic wordnet was undertaken by Diab [45]. She introduced an automated approach known as SALAAM (Sense Assignment Leveraging Annotations And Multilinguality) to translate synsets from PWN into Standard Arabic. This translation process relied on PWN 1.7 and an English-Arabic corpus as knowledge sources. Notably, her primary focus was on translating lemmas without glosses and example sentences. This approach was evaluated using a dataset comprising 447 synsets.

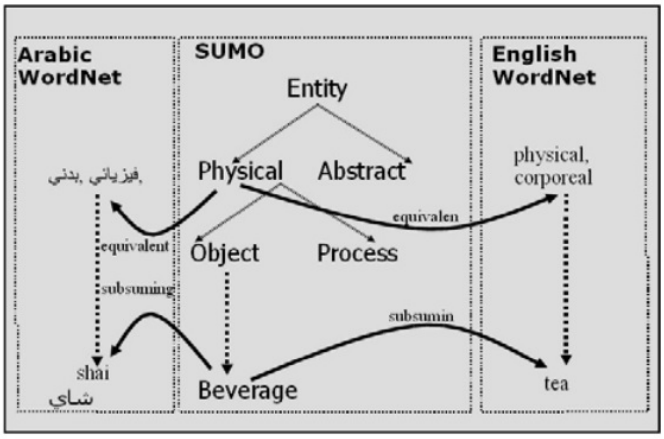


Figure 5.1: SUMO mapping to WordNets [51]

AWN V1 represents the inaugural Arabic WordNet developed by Elkateb et al. [51]. The development approach closely mirrors the methodology employed in creating EuroWordnet [154], which consists of two phases. The first phase involves constructing a foundational core wordnet centered around *base concepts* [153], while the second phase focuses on expanding the core wordnet's coverage by incorporating additional criteria. This version of AWN is aligned with PWN in terms of structure and content covering WordNet domains defined by Magnini et al. [101]. This wordnet also integrates the Suggested Upper Merged Ontology (SUMO) to provide a formal semantic framework [51].

In the case of the core Arabic WordNet, the process involves encoding the Common Base Concepts (CBCs) found in the EuroWordnet and BalkaNet [148] as synsets. This is achieved through a manual translation effort, wherein all English synsets having an equivalence relation in the SUMO ontology are translated into their corresponding Arabic synsets. Figure 5.1 illustrates this process, showing an example of how Arabic Synsets are linked to overarching SUMO terms that directly correspond to the associated English synsets. Each translated synset is validated by evaluating the coverage of synset lemmas and the domain distribution of these synsets. These efforts produced 9,228 synsets in the core wordnet of AWN V1. The distribution of these synsets, categorized by Part-Of-Speech (POS), is detailed in Table 5.8.

To expand the core of AWN, Elkateb et al. [51] introduced the Suggested Translation semi-automatic method, using available bilingual (Arabic-English) resources to extract *<English word, Arabic word, POS>* tuples. This method served a similar purpose in the development of Spanish WordNet [53] and BalkaNet [148]. Building on eight heuristic procedures, associations between Arabic words and PWN synsets were assigned scores, relying on Arabic-English bilingual resources. Lexicographers utilized these scores to create

POS/WN	AWN V1 (Core WN)	AWN V1 (Ext. WN)	AWN V2
Noun	6,252	6,558	7,960
Verb	2,260	2,507	2,538
Adjective	606	446	271
Adverb	106	107	500
Total	9,228	9,618	11,269

Table 5.8: The count of Arabic synsets in each AWN version based on POS

new synsets or supplement existing ones with additional lemmas. The total number of synsets in this version is 9,618.

After the first release of AWN, there were many attempts to enrich its content concerning the number of synsets, lemma, and the relations between them. Alkhalifa et al. [6] introduced an automated method to enhance the coverage of named entities (NE) within AWN V1. This method used Wikipedia and established connections to PWN 2.0. In this study, 1,147 synsets were generated, covering 1,659 NEs across 31 general categories. In these studies, Boudabous et al. [31] and Batita et al. [18] proposed a hybrid linguistic approach grounded in morphological patterns. They used Wikipedia and PWN to enrich AWN with new semantic relations. The former augmented AWN by establishing relations between nominal synsets, while the latter incorporated antonym relations.

As part of the ongoing efforts to enrich AWN, Abouenour et al. [1] introduced a semi-automatic method to increase the coverage of AWN V1. Their objective was to enhance NEs, verbs, and noun synsets. For the enrichment of NE synsets, the authors present a three-step methodology, which translates YAGO (Yet Another Great Ontology) entities [142] into Arabic instances and extracts Arabic synsets. Regarding verb synsets, the authors adopted a two-step approach inspired by Rodríguez et al. [129]. The first step involved suggesting new verbs by translating a set of verbs from VerbNet [136] into standard Arabic. In the second step, Arabic verbs were interconnected with AWN synsets by establishing a graph connecting each Arabic verb with its corresponding English verbs in PWN. The authors employ a two-step method that detects hyponym/hypernym pairs from the web to enrich noun synsets. The overall result of this work is introducing a new version of AWN, known as AWN V2, including 11,269 synsets (for more details, see Table 5.8).

Despite previous efforts focused primarily on expanding the coverage of synset lemmas, AWN still falls short compared to other WordNets in terms of content quality. This limitation was highlighted by Batita et al. [18],

who emphasized in their research that "*AWN has very poor content in both quantity and quality levels*". Our work addresses the synset quality level, focusing on the dimensions of synset correctness and completeness.

AWN V1 represents a significant milestone for several reasons. First, it includes the most common concepts and word senses found in PWN 2.0, ensuring broad representation in AWN V1. Second, its design and integration with PWN synsets facilitate cross-language usability. Finally, like other WordNets, AWN establishes a connection with SUMO, further enhancing its utility. However, several issues related to synset quality have been identified in most synsets within this resource. These issues are also present in AWN V2, as follows: (1) all synsets lack gloss and/or illustrative examples; (2) many synsets contain incorrect senses, lemmas (including incorrect word forms or repeated words), and incorrect relations between synsets; (3) many synsets lack essential senses, lemmas, and necessary relations. For instance, consider the following synset {ضجيج، ضوضاء} presented in AWN V1, corresponding to the English synset "*noise: sound of any kind, especially unintelligible or dissonant sound; he enjoyed the street noises*". In AWN V2, this synset was enriched to include {ضجج، ضوض}, resulting in {ضجيج، ضوضاء، ضجج، ضوض}. In this case, the synset incorporates two erroneous lemmas {ضجج، ضوض}, which are not found in Arabic dictionaries such as the Almaany dictionary. Additionally, it lacks the lemma ضجة, which means "*noise*".

For these reasons, AWN V1 was integrated with the *Arabic lexicon* in the UKC. Similarly, in this thesis, we selected AWN V1 over other Arabic WordNet versions for enhancement, focusing on addressing untranslatable cases, correcting erroneous lemmas, and expanding synset coverage by adding missing components.

5.5.2 Experiment

This section demonstrates the use of the expert-driven methodology described in Chapter 4 in improving the content quality of the Arabic wordnet, by considering the following goals in this experiment:

1. Language diversity and phrasets: As introduced in Section 3.2, ideas are expressed in cultures in different ways, which leads to lexical gaps in some languages. Another phenomenon in Arabic (and maybe in other languages) is the usage of prepositional phrasets to express a synset meaning. For example, the meaning of this synset "*someday: some unspecified time in the future; someday you will understand my actions*" is identified as a lexical gap in Arabic, and the phraset يوماً ما is used to express this meaning. We add these phrasets to the Arabic WordNet to increase the understandability of synsets. Also,

such phrasets can be used in NLP applications to identify the intended synset.

2. Synset glosses: Each synset should have a gloss that clearly identifies its meaning. Without such gloss, we will not be able to understand the synset, moreover, we will not be able to differentiate between the meanings of the same lemma in different synsets, for example, the word "*love*" has more than one meaning e.g, belongs to different synsets

3. Synset examples: Each synset should have at least one example to clarify its usage. Such examples also allow us to verify the synonymity between the synset lemmas. This is crucial for the synset correctness.

The process of AWN improvement follows the expert-driven methodology with two modifications. First, during the contribution task generation, instead of collecting language-independent concepts from the supra-lingual concept layer of the UKC, we collected synsets (see Table 5.9) from the Arabic lexicon—rooted in AWN V1—and their corresponding English synsets from the English lexicon. These synsets (Arabic and English) were compiled into a spreadsheet for translation. The synsets were organized into four categories (each on a separate sheet) based on their part of speech (POS). Each row in the spreadsheet represents a synset and includes details such as the synset ID, lemmas, gloss, and example sentences in both Standard Arabic and English. Additional empty slots were provided to insert missing lemmas, glosses, examples, and comments by the translator. One additional slot was designated for validation purposes and included a space for comments from the validator. In this step, the linguistic expert excludes all (42 synsets) named entities from the spreadsheet.

Second, during the contribution collection phase, instead of focusing on identifying equivalent words and lexical gaps, the translator added phrasets to improve the understandability of synsets, inserted missing lemmas and example sentences, and corrected errors (e.g., removing incorrect lemmas) in the original Arabic synsets. Further details on the contribution collection process in this experiment are provided below.

	Noun	**Verb**	**Adjective**	**Adverb**	**Total**
Synsets	6,516	2,507	446	107	**9,576**
Words	13,659	5,878	761	262	**20,560**

Table 5.9: The count of synsets and words (imported from the Arabic lexicon) in the input dataset based on POS

Contribution collection

The process of synset enhancement was carried out by a translator, who is an Arabic native speaker. Regarding his socio-linguistic background, he possesses a bachelor's degree in the field of translation (English-Arabic). Before the translation, in addition to providing him with instructions presented in Section 4.2, the translator has been trained as described in the following subsections.

Synset understanding: Central to this process is ensuring that the translator possesses a clear understanding of the synset he is tasked with translating. Misunderstandings can arise when the translator does not grasp a thorough understanding of both the synset lemmas and the gloss in English. The translator is asked to understand each English synset in the spreadsheet using the following notable instructions: (1) Use external resources such as dictionaries and Wikipedia to understand the meaning of the synset; and (2) Given the authority to skip the synset or leave a comment when he does not understand the meaning of the synset.

Lexical gap identification and synset lexicalisation: In this step, for each English synset in the spreadsheet, the translator decides whether it has an equivalent meaning in Arabic (lexicalisation exists) or is a lexical gap based on the understanding of the English synset and using a bilingual dictionary. If an English synset$_i$ is a gap, the translator performs **step A**; otherwise, he performs **steps B** and **C**.

Step A: Lexical gap processing, in this step, the translator is asked to mark the English synset$_i$ as a lexical gap in the spreadsheet and provide a phraset in Arabic. For example, the synset "*expressively: with expression, in an expressive manner; she gave the order to the waiter, using her hands very expressively*" is identified as a lexical gap in Arabic, and بشكل معبر meaning "*an expressive way*" is provided as a phraset to this synset.

Step B: Synset translation, after the translator confirms the existence of the equivalent meaning of the English synset$_i$ in Arabic, he translates this synset to Arabic. This translation includes the following steps.

1. **Translating synset gloss:** The translation is across language and cross-cultural communication. It should give a complete transcript of the synset; meanwhile, the style and manner of writing should be at least the same quality as the gloss of English. Above all, faithfulness, expressiveness, and closeness are the important three elements of translation. The gloss should explicitly express the semantics and the common attributes of a synset.

2. **Translating synset lemmas:** The translator should keep two key considerations in mind while translating synset lemmas. Firstly, this translation process does not entail a direct one-to-one correspondence between English and Arabic terms. Secondly, it is important to note

that the set of lemmas within the English synset may not be exhaustive, meaning that it might not contain all the synonyms associated with the synset. To translate the synset lemmas, we go through the following phases:

- **English lemmas translation:** Translate the English synset lemmas into Arabic. The result of this step is a set of lemmas of the length n, where n is the number of lemmas in the English synset.
- **Arabic synonyms collection:** For each translated lemma, the translator collects the lemma synonyms in Arabic. The result of this phase is m synonym sets in Arabic, $m \leq n$ (since some Arabic lemmas may have empty synonym sets).
- **Arabic synonyms validation:** Based on the synset gloss, for each of the m synonym sets in Arabic, the translator excludes all synonyms that do not belong to the synset. Use the provided examples in the English synset and other examples to include/exclude the synonyms in this phase.
- **Arabic lemmas collection:** The translator collects the Arabic lemmas, resulting in the translation process in phase (1) and the synonyms produced from phase (2) and put them as the Arabic synset lemmas.
- **Arabic lemmas ordering:** The translator orders the Arabic collected lemmas in phase (4), wherein the first lemma is the Arabic synset preferred term and so on (in descending order of importance). Based on the examples provided in phase (3) (and other examples if needed), the translator gives preferences for the lemmas based on these examples.

3. **Translating synset examples:** Examples within a synset contribute to a clearer comprehension of how to utilize the synset lemmas, consequently enhancing the overall understanding of the fully lexicalized synset. We employ the same examples crafted during the lemma translation phase as synset examples. This approach signifies that we do not solely translate the examples found in the English synset. Ideally, we provide an example sentence for each Arabic synset, aiming to include examples for every synonym within the synset, even when the corresponding English synset lacks examples entirely. These examples are integrated into the Arabic synsets, aligning them with the order of their respective synonyms.

Step C: Comparing the produced (translated) synset in Step (B) with the corresponding synset from the Arabic lexicon, at this stage,

the translator compares the translated synset generated in Step B and its corresponding original Arabic synset, as imported from the Arabic lexicon. This Arabic synset is designated to correspond with the English synset$_i$ in the spreadsheet. Based on the gloss and examples provided in the generated synset, the translators undertake the following actions: (1) Copy lemmas from the translated synset to the original Arabic synset if they are missing from the original synset. (2) Exclude lemmas from the original Arabic synset, which are not covered by the synset gloss and examples. (3) Copy the gloss and the examples from the translated synset to the original synset if they are missing in the latter.

For validation, a linguistic expert employs the lexicon-level validation methodology described in Section 4.3 to verify the produced synsets, gaps, and phrasets. The validation of synset components involves three steps: (1) lemma validation: the synset lemmas must be accurate and complete; (2) example validation: each lemma should include at least one example that illustrates its intended usage; (3) gloss validation: the Arabic gloss must be clear, free of typographical or grammatical errors, and accurately convey the intended meaning of the English synset. Unclear words, as well as non-confirmed gaps or borderline cases that are neither clearly gaps nor non-gaps, are reviewed and confirmed by the lexical-semantic expert responsible for managing the UKC system.

In this study, four experiments (one for each part of speech) were conducted to evaluate AWN V1 synsets and address synset quality issues. The experiments involved 6,516 nouns, 2,507 verbs, 446 adjectives, and 107 adverbs. In each experiment, for every Arabic synset listed in the spreadsheet, the translator was tasked with translating the corresponding PWN synset into Arabic or identifying it as a lexical gap using a bilingual (English-Arabic) resource, such as the Al-Mawrid Al-Qareeb المورد القريب dictionary [12]. If a lexicalization existed in Arabic, the translator addressed this by comparing the translated Arabic synset with the original Arabic synset (AWN V1 synset) in the same row. This process involved adding missing synset lemmas, glosses, and example sentences and/or correcting incorrect elements. Additionally, if the English synset represented a lexical gap in Arabic, it was marked as such, and a phraset was provided to express the synset. (Note that phrasets were also used for some translated synsets to enhance understandability).

Validation was carried out by an Arabic linguistic expert who has a Ph.D. in the Arabic language and is a university instructor at the linguistics department. Results can be seen in Table 5.10, where by correctness we understand the number of contributions validated as correct divided by the total number of contributions. These contributions can be newly added or deleted lemmas, collected glosses and example sentences, identified lexical gaps, and inserted phrasets. For example, in the case of an added lemma, the validator either confirms the addition or rejects it by leaving a comment. For

Contribution	Correctness
New lemmas	97.34%
Deleted lemmas	98.89%
New glosses	98.76%
New examples	99.13%
Gaps	96.82%
Phrasets	97.54%
Total	98.08%

Table 5.10: Validator evaluation of translator contributions

instance, مِقْيَاس meaning "*a measuring tool*" is deemed an incorrect added word to the synset {مِقْدَار، قَدْر، كَمّ، كَمِّيَّة} which corresponds to *{measure, quantity, amount: how much there is of something that you can quantify; he has a big amount of money}*. In the case of identified gaps, the validator either as confirmed gaps or as non-gaps due to an existing lexicalization in Arabic, which the validator needs to indicate. For instance, the following English synset *{try, try on: put on a garment in order to see whether it fits and looks nice; Try on this sweater to see how it looks}* is considered a gap. The validator rejected it and provided this word قَاسَ with the same meaning.

Upon discussion between the validator (linguistic expert) and the translators, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts provided in English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

The experimental results demonstrate that the effort to collect contributions resulted in the update of 5,554 out of 9,576 synsets in the Arabic lexicon. This includes the identification of 236 lexical gaps and the addition of 701 phrasets. Further details on these results are provided in Section 10.1.5. The new version of the Arabic lexicon is publicly available as the third version of the Arabic WordNet (AWN V3). To our knowledge, AWN V3 is the first Arabic WordNet to include lexical gaps and provide corresponding phrasets.

Part III

Crowdsourcing Lexical Diversity

6 Crowdsourcing

Contents

Crowdsourcing has become a widely adopted approach in NLP and computational linguistics for data collection, annotation, and evaluation. In the context of building diversity-aware resources, crowdsourcing not only accelerates the acquisition of comprehensive lexical data but also enhances the cultural sensitivity of semantic representations by incorporating a broader range of societal perspectives through a diverse pool of contributors, typically recruited from online platforms.

This chapter provides an overview of crowdsourcing, beginning with Section 6.1, which introduces a formal definition of crowdsourcing, explores key areas of its application, describes the general structure of crowdsourcing systems, and reviews some of the most commonly used crowdsourcing platforms and their purposes. In Section 6.2, we discuss the need for crowdsourcing in building diversity-aware multilingual lexicons. Finally, Section 6.3 reviews the state-of-the-art use of crowdsourcing in developing language datasets.

6.1 Crowdsourcing: key definitions, applications and process flow

Crowdsourcing is a collaborative approach that harnesses the collective efforts of a large, diverse group to solve problems, generate ideas, or complete specific tasks. Coined by Howe [73], the term combines *crowd* and *outsourcing* referring to the process of distributing tasks or problems to a dispersed, often online, crowd. Crowdsourcing relies on voluntary participation and can include a variety of tasks, from data collection and problem-solving to content creation and innovation.

Estellés-Arolas et al. [52] introduced an integrated definition for crowdsourcing, describing it as a participative online activity where individuals or groups propose tasks through open, flexible calls. Participants vary widely in knowledge, skills, and demographics, making crowdsourcing a versatile approach that leverages human capital beyond traditional organizational boundaries.

The crowdsourcing paradigm has seen rapid adoption across multiple domains, providing solutions that are cost-effective, scalable, and driven by open collaboration. Key areas of crowdsourcing applications include:

1. *Business and marketing*: Crowdsourcing supports product development, market research, and customer feedback. Companies often solicit ideas from consumers to improve products or services, thereby creating campaigns more targeted and innovative [29].

2. *Health and medicine*: Crowdsourcing aids medical research, diagnosis, and public health initiatives. Platforms like PatientsLikeMe [138] enable patients to share experiences, helping others with similar conditions while contributing to a larger pool of health data for researchers.

3. *Education and learning*: Educational platforms like Duolingo [145] and Khan Academy [109] employ crowdsourcing for content development, translation, and community support. Peer-to-peer learning and collective feedback enhance accessibility and the quality of learning resources.

4. *Science and research*: Known as *citizen science*, crowdsourcing allows non-experts to contribute to data collection and analysis. Projects like Galaxy Zoo [57] involve the public in tasks that require human pattern recognition, helping researchers manage large data volumes.

5. *Software development*: Crowdsourcing supports rapid prototyping, bug fixing, and quality assurance in software engineering, offering cost-effective solutions and diverse expertise[1].

6. *Image and video labeling* [90]: Crowdsourcing facilitates large-scale data annotation for machine learning, including image classification and object recognition, essential for training AI models.

7. *Natural language processing (NLP)*: Tasks like text transcription, language translation, and sentiment analysis are often crowdsourced to capture linguistic nuances beyond automated capabilities [74].

As illustrated in Figure 6.1, a crowdsourcing system typically consists of four main components: a task, a task requester, crowd, and a digital platform that facilitates interaction [29]. The *task* is the objective or project to

[1] https://www.topcoder.com/

be completed, often broken down into smaller, manageable units for crowd processing. The *task requester*—often an individual, organization, or company—defines the task and submits it to the platform, which disseminates it to the crowd. The *crowd*, or *contributors*, participates by offering ideas, providing solutions, or completing tasks as specified. The *platform* facilitates task distribution, communication, and, in some cases, the allocation of rewards based on contributors' performance or task requirements. The trustworthiness or reliability of the crowd's past contributions or performance is a key factor in determining their *reputation*.

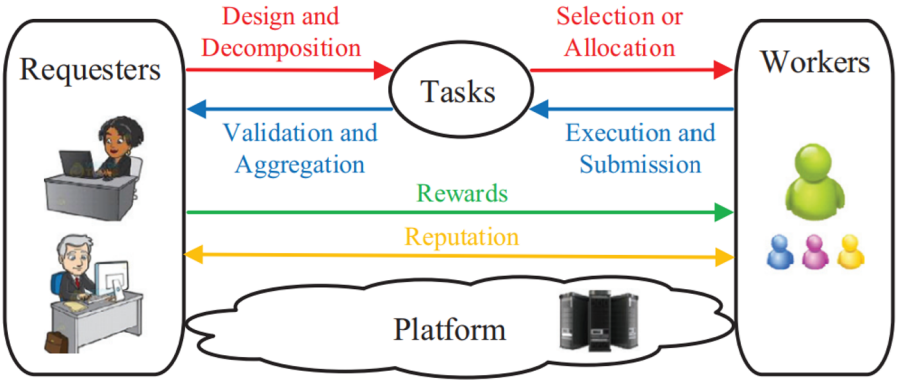


Figure 6.1: Components and processes in crowdsourcing systems [29].

Crowdsourcing platforms have become essential tools for tapping into the collective intelligence of large groups, enabling various industries and researchers to gather data, generate insights, and complete tasks efficiently. Some prominent crowdsourcing platforms used in research, technology development, and data collection include:

1. *Amazon Mechanical Turk (MTurk)*[2], one of the most widely used crowdsourcing platforms, provides access to a large pool of microtask workers. It allows requesters to post small tasks, often requiring basic data processing, surveys, or classification, which individual workers can complete relatively quickly. MTurk is frequently used in social science and behavioral research for survey distribution, data labeling for machine learning, and gathering human-verified responses to tasks that machines alone struggle to perform accurately.

2. *CrowdFlower/Appen*[3]: Originally known as CrowdFlower and now rebranded as Appen, this platform specializes in data annotation and labeling, especially for training AI and machine learning (ML) models. The platform integrates human insight to improve data quality and

[2] https://www.mturk.com/
[3] https://www.appen.com/

enhance training datasets. It is often used for tasks like image labeling, transcription, and sentiment analysis, where high-quality labeled data is crucial for training algorithms in NLP or computer vision.

3. *Zooniverse*[4] is a crowdsourcing platform tailored for research projects, especially in the fields of natural sciences, humanities, and social sciences. It enables volunteers to assist with tasks such as data analysis, including identifying patterns in images or transcribing historical documents. Zooniverse has hosted a wide range of projects, from classifying galaxies and animals in camera trap photos to transcribing ancient manuscripts. Relying on citizen scientists, it has become popular for public outreach and engagement.

4. *Upwork*[5] is a general freelance marketplace, often used for more complex and larger-scale tasks that require specialized skills. Unlike microtask platforms, it allows requesters to post jobs that can be completed by freelancers with specific qualifications. The platform is commonly used for tasks that require a higher level of expertise, such as software development, data analysis, and report writing, as well as tasks that need consistent freelancer involvement over time.

5. *InnoCentive*[6] is a crowdsourcing platform geared towards problem-solving, where companies or researchers can post challenges that require innovative solutions. Solvers from around the world can submit their ideas, and the best solutions often receive monetary rewards. This platform is widely used in innovation-focused sectors such as pharmaceuticals, engineering, and environmental sciences, where problems are complex and require a diversity of ideas to achieve breakthrough solutions.

6. *Kaggle*[7], a subsidiary of Google, is a platform for data science competitions where individuals or teams compete to produce the most accurate models or analyses for particular datasets. Companies or research institutions can post datasets for specific goals, and participants work to develop algorithms or predictive models. Kaggle is popular for competitions in predictive modeling, machine learning, and data analysis.

7. *Prolific*[8] is a crowdsourcing platform designed specifically for academic and scientific research. It connects researchers with a diverse, prescreened participant pool from over 38 countries, ensuring reliable

[4] https://www.zooniverse.org/
[5] https://www.upwork.com/
[6] https://www.wazokucrowd.com/
[7] https://www.kaggle.com/
[8] https://www.prolific.com/

data collection across various study designs, including surveys, experiments, and longitudinal studies. Prolific places a strong emphasis on participant quality, ethical standards, and fair compensation, making it particularly suitable for researchers seeking high-quality data with strict adherence to ethical guidelines.

6.2 The need for crowdsourcing in our research

In constructing LSRs that reflect the diverse and rich nature of multilingual contexts, crowdsourcing has emerged as a powerful solution for capturing nuances specific to language and culture, which are integral to creating truly inclusive and representative resources. Crowdsourcing allows researchers and developers to leverage the collective knowledge of a diverse array of contributors, resulting in several advantages:

1. *Wide reach*: Crowdsourcing provides access to an *extensive network of contributors* from various linguistic and cultural backgrounds. By engaging people from around the world, researchers can gather diverse perspectives, *local expressions*, and *variations in language use* that would be challenging to capture otherwise. This diversity enriches the lexical-semantic resource with insights that reflect real-world linguistic variation and help prevent cultural biases that may arise from using a homogenous set of contributors [133].

2. *Up-to-date language use*: Language evolves rapidly, with *new words*, *slang*, and *expressions* emerging continually. Crowdsourcing enables researchers to capture these linguistic shifts in real-time by engaging contributors who use and understand current language trends [43]. This approach helps in maintaining the relevance and modernity of lexical resources, ensuring they reflect *contemporary language* as it is used in everyday settings.

3. *Inclusivity and representation of minority kanguages*: Traditional linguistic resources often overlook minority languages, dialects, and sociolects, which can lead to an incomplete or skewed representation of language diversity. Crowdsourcing helps bridge this gap by inviting contributions from speakers of *underrepresented languages* and dialects. This inclusivity supports the creation of more comprehensive resources, capturing the unique linguistic characteristics of diverse communities and promoting linguistic equity [110].

4. *Scalability*: Developing extensive and detailed linguistic databases typically requires substantial data collection. Crowdsourcing offers a scalable solution, allowing researchers to quickly gather *vast amounts*

of data (compared to traditional methods) by engaging large numbers of contributors simultaneously [140]. This scalability is especially valuable in multilingual projects, where building large cross-linguistic databases is essential for accurate modeling and representation.

5. *Contextual and local insights*: Native speakers bring invaluable *context-specific knowledge*, offering *authentic examples* of language use that reflect cultural and situational nuances. Crowdsourcing taps into this local expertise, enabling resources to better capture and represent language subtleties in specific contexts [162]. Such contributions are crucial for improving the accuracy and relevance of LSRs, as they are grounded in the authentic language experiences of native speakers.

6. *Cost-effectiveness*: Building large-scale multilingual resources typically involves significant costs, as it requires expertise across multiple languages and dialects. Crowdsourcing provides a more affordable alternative by leveraging *voluntary contributions* from a global network of participants, thereby reducing the need for extensive hiring or formal data collection efforts [150]. This cost-effective approach is particularly beneficial for *low-resource languages*, which often lack funding and expertise for formal data collection.

The crowdsourcing approach supports the creation of a more accurate, diverse, and comprehensive linguistic resource, which is essential for applications in NLP and cultural preservation in a globalized world. By harnessing the strengths of crowdsourcing, researchers can compile LSRs that are not only rich in linguistic diversity but also aligned with contemporary language trends, inclusive of minority and indigenous languages, and reflective of authentic, real-world language usage.

6.3 State of the art: crowdsourcing language data

Crowdsourcing has been used to create a wide range of linguistic resources, including lexical-semantic data. Ganbold et al. [61] used translation to create a WordNet in Mongolian and proposed a two-phase crowdsourcing workflow for localizing WordNet synsets into less-resourced languages, such as Mongolian. The approach involves a translation phase, during which crowd workers (volunteers) recruited via the CrowdCrafting[9] platform suggest synonymous words by translating English PWN synsets into Mongolian. This is followed by a validation phase that employs statistical inter-rater agreement metrics (Fleiss' kappa and Krippendorff's alpha) to ensure quality. This method achieved a precision of 0.74 for 947 synsets, demonstrating the efficiency of the crowdsourcing approach. Wijesiri et al. [161] also employed the

[9] https://crowdcrafting.org/

translation method to bootstrap the Sinhala wordnet from English to Sinhala, with help from internet users proficient in both languages, using PWN as the source. Lanser et al. [92] used the translation approach to create a Japanese lexicon from DBpedia. They started with building an English lexicon for the DBpedia ontology and recruited annotators on CrowdFlower to translate it into Japanese.

Benjamin et al. [27] presented a crowdsourcing model enabling the development of lexicons in low-resourced languages from scratch, utilizing the "*Fidget Widget*" mobile application with internet users' assistance. Biemann et al. [30] showcased a method for generating a sense inventory from scratch, employing annotators via Amazon Mechanical Turk (MTurk) to perform word sense acquisition tasks for English. El-Haj et al. [70] recruited annotators on MTurk to construct the Essex Arabic Summaries Corpus (EASC), resulting in 2,360 sentences and 41,493 words in Jordanian and Gulf Arabic.

Manerkar et al. [103] created the "*Konkani Shabdarth*" crowdsourcing website to enhance the Konkani wordnet's content quality. The platform, hosted on the Goa University website, offers two roles: players, including Konkani speakers like students and faculty members, who add missing words to synsets, and experts, such as aspiring authors and writers, who validate synsets. SloWCrowd, a crowdsourcing platform introduced by Fier et al. [56], addresses error correction in the Slovene wordnet [62] using PWN synsets as a reference. Cibej et al. [40] conducted crowdsourcing tasks to clean the Thesaurus of Modern Slovene, recruiting six workers on PyBossa[10] to judge the usefulness of synonym pairs. Chandran Nair [37] presented a crowdsourcing method to enhance Malayalam wordnet using Google Forms integrated into a mobile application, referencing PWN.

We are aware of only one prior attempt to collect lexical gaps or evidence of lexical untranslatability using crowdworkers. Giunchiglia et al. [68] introduced a crowdsourcing tool for translating English lexicalizations into Italian, including identifying lexical gaps, by recruiting linguistic experts as the crowdworkers. Our work differs significantly from this approach as it utilizes crowdsourcing with ordinary native speakers, rather than expert-sourcing. Additional differences between our approach and the previous efforts can be summarized as follows:

1. *No reliance on English as an intermediary*: Our crowdsourcing methodology does not depend on English as a mediating language. Instead, we compare datasets from any two languages within the same semantic domain, thereby avoiding the pervasive issue of *English bias*.

2. *Focus on untranslatability and bidirectional exploration*: We focus exclusively on identifying instances of untranslatability by enabling *bidirectional* exploration of lexical gaps between language pairs, without

[10] https://pybossa.com/

assuming a unidirectional translation process. This approach leverages contributions from native speakers through our crowdsourcing methodology.

3. *Incorporation of state-of-the-art quality control measures*: Our approach ensures data quality and reliability by *selecting skilled annotators* based on inter-rater agreement. Crowdsourced data is validated through a novel *inter-rater agreement-based mechanism*. Additionally, we employ quality control measures such as *Attention Check Questions (ACQs)* and *completion time* metrics.

7

Crowdsourcing Methodology

Contents

This chapter establishes a methodology for the crowd-based collection of evidence on lexical diversity. Diversity is inherently an interlingual notion, referring to phenomena not shared across languages. Such phenomena are particularly frequent in, but are not limited to, socially or culturally important domains: food, religion, family relationships [84], motion verbs [155], body parts [159], colors [104], spatial dimensions [91], food [11], cutting and breaking events [102], pain predicates [126], perception verbs [151], or putting and taking events [87]. The initial inputs of our methodology are, therefore, 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 a difference: for the language pair (A, B), it will find lexicalizations present in A but not in B, and for (B, A) the opposite: lexicalizations present in B but not in A.

The methodology is divided into three main steps: (1) task generation; (2) crowdsourcing; and (3) validation. *Task generation* is a semi-automated step that produces two lists of *lexical entries*, one in each language. A lexical entry is a tuple (*word*, *gloss*), i.e., a term and its definition.

A *crowdsourcing micro-task* consists of one lexical entry from the source language (SL), as well as all lexical entries from the target language (TL) as candidates. The goal of crowdsourcing is to indicate whether the SL entry is translatable to the TL and, in the positive case, to provide the equivalent TL entry. Our policy is that crowd workers need to be native speakers of the TL as well as having a sufficient command of the SL. By "sufficient command" we mean the ability to understand the precise meaning of the SL lexical entry, potentially with the help of lexicons or the Internet. Deciding whether the term has an equivalent in the TL, on the other hand, is a more difficult task requiring a deep familiarity and an active knowledge of the

language, which is why we require workers to be native speakers of the TL.

The *validation* step integrates worker contributions through an automatic validation process, followed by a final expert-based verification. The output of the method is a list of words in the SL that are annotated either as untranslatable or as lexicalized in the TL, with the equivalent word provided in the latter case. A general overview of the crowdsourcing methodology is depicted in Figure 7.1.

In this methodology, two key roles are defined: 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 the TL within the SF, (2) designing and creating the crowdsourcing tasks, (3) overseeing task execution to ensure the quality of contributions, and (4) validating and exporting the finalized crowdsourced data. On the other hand, the crowd worker's role focuses on contributing to the task by identifying equivalent terms and lexical gaps in the TL.

To ensure the quality of the crowdsourced data, we implemented quality control mechanisms across three phases. The first phase includes *pre-crowdsourcing quality control*, which selects high-quality crowd workers using two tools: (1) a proficiency test that assesses contributors' abilities in performing crowdsourcing tasks, and (2) Crowd Filtering using Alpha, an algorithm designed to identify low-quality workers—those who answer test questions randomly—from the pool of proficient workers. This is achieved by measuring inter-annotator agreement (IAA) using Krippendorff's Alpha.

The second phase employs *live quality control* during the crowdsourcing process, utilizing the following mechanisms: (1) Attention Check Questions (ACQs), which are straightforward questions related to task content designed to assess worker attentiveness and filter out those who may be rushing through the task; and (2) completion time monitoring, which tracks how long a contributor takes to complete each question. Significant deviations from a worker's average completion time are treated as potential indicators of low-quality responses.

Finally, the third phase involves *post-crowdsourcing quality control*, consisting of two sequential validation mechanisms: (1) Crowdsourced Data Validation using Alpha, an automatic process that evaluates IAA among workers' responses using Krippendorff's Alpha, filtering out answers that fall below a predetermined threshold; and (2) expert validation, in which a linguistic expert reviews responses that did not reach 100% IAA but exceeded the minimum threshold.

7.1 Step 1: task generation

Task generation takes lexical resources and corpora as input and produces a list of micro-tasks (*word–gloss* tuples) for the workers. Micro-task generation

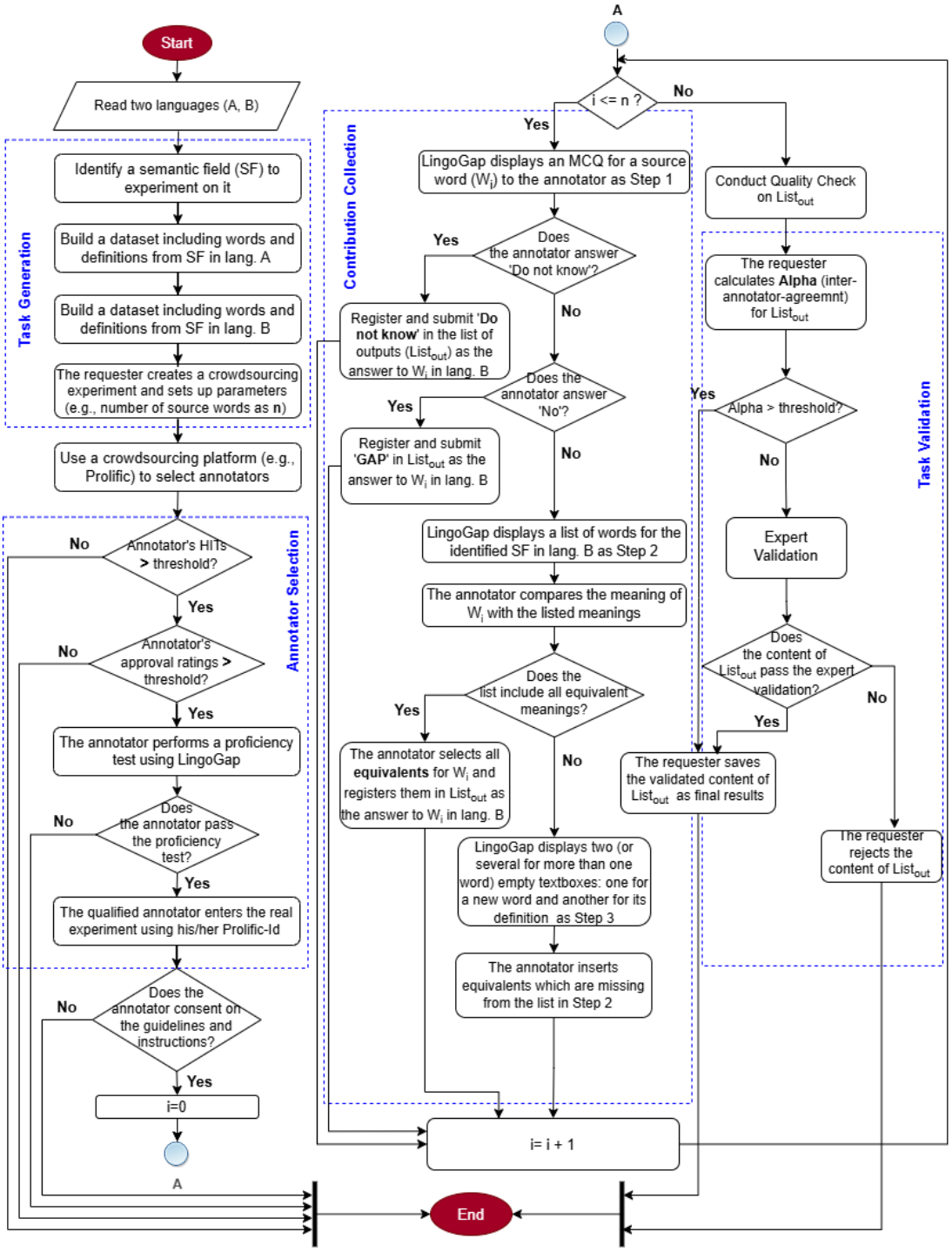


Figure 7.1: Overview of Crowdsourcing Methodology.

is semi-automatic: we algorithmically produce candidate micro-tasks that are subsequently filtered by an expert. The challenge lies in identifying candidate words that belong to the *semantic field*, in a way that is robust enough to be applied to a wide range of languages and dialects, including low-resourced ones.

We achieve this robustness by combining a variety of language resources, depending on what is available for the given language and semantic field:

- mono- or multilingual lexical databases, wordnets;
- online dictionaries and encyclopedias (e.g. Wikipedia, Wiktionary, traditional monolingual dictionaries);
- language models or word embeddings;
- digital or undigitized corpora;
- field work with native speakers.

Lexical databases, such as the UKC or other multilingual or monolingual wordnets, are our preferred source of data as they contain rich, structured, meaning-annotated lexical entries that also contain definitions. Domain or topic annotations as well as broader–narrower hiearchies in lexical databases can be used to filter entries relating to the semantic field. However, only a handful languages in the world are covered by large-scale, high-quality lexical databases, and those that systematically include glosses are even more rare.

Online dictionaries and encyclopedias are used in two ways: (1) they can provide glosses whenever these are not directly available from the lexical database. Given a candidate word, we retrieve the first sentence from Wiktionary or Wikipedia and use it as a candidate gloss. If no entry for the word is found then it is discarded (no task is generated). (2) They can provide words that may or may not belong to the semantic field, to be filtered using vector similarity methods described below.

Language models and word embeddings are used to generate lexical entries if no lexical resource is available in the source language for the given semantic field, or if its coverage is not sufficient. The idea is to traverse an existing word list, obtained from a dictionary as described above, to identify those that belong to the given semantic field in the given language: e.g., *cake*, *bread*, or *tomato* belong to the field of *food*. From a small text corpus describing a semantic field, consisting of a definition and a set of example terms, a vectorial representation is computed: sentence vectors from a language model (e.g., AraBERT for Arabic) or word vectors from a word

embedding. These sentence or word vectors are then compared to the vectors of dictionary words using cosine similarity, and the most similar ones are retained.

Text corpora can be used to train word embeddings or language models whenever they are not available for a language. For languages where no digital corpora are available, an initial corpus digitization step is necessary.

Field work is an often neglected yet efficient method for obtaining a high-quality, focused corpus of words and definitions for a given language. This method can be used for low-resource languages and dialects when a direct contact with native speakers is available. First, a language expert provides an initial set of seed words that belong to the semantic field. Then, native speakers provide words and definitions related to the field and the seed words.

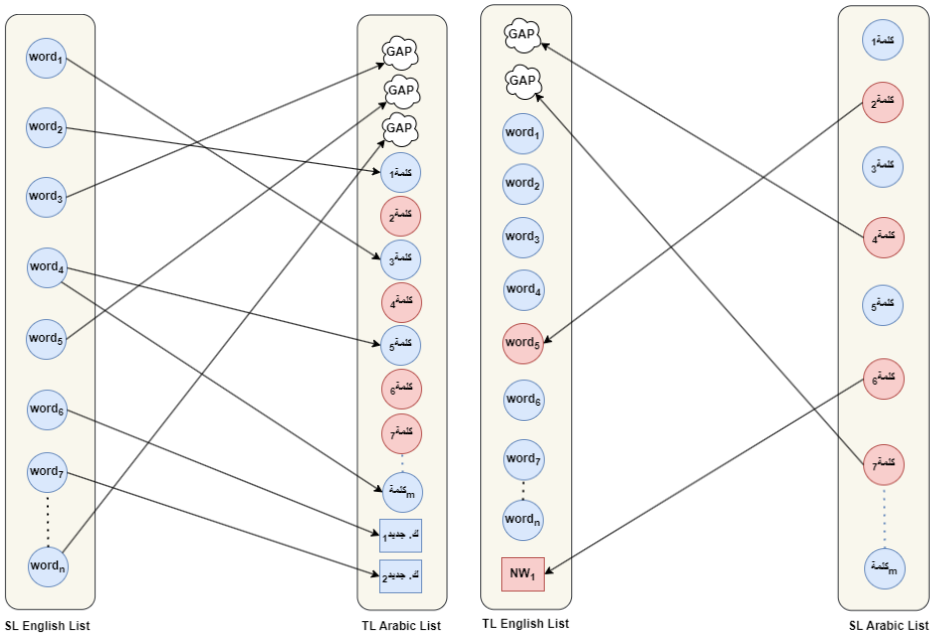
For the most widely spoken languages, such as English, Spanish, or Chinese, task generation is possible using a lexical database alone, such as the UKC, but any combination is applicable from the above resources and methods. For languages where lexical databases are of lower quality, have an inadequate coverage, or do not exist, such as Arabic or Hungarian, data from lexical databases can be extended by entries obtained from traditional lexicons, filtered using language models. For even lower-resourced languages, which do not have language models but still possess usable-sized corpora, such as European minority languages and dialects, digital dictionaries can provide input words, and word embeddings can be trained and used for filtering according to the semantic field. Finally, in the case of dialects and severely endangered languages with few or no existing corpora, results from preexisting field work—such as kinship terms for Arabic dialects [80]—can be used to obtain smaller-scale but potentially good-quality word lists.

7.2 Step 2: crowdsourcing

This section explains how the requester uses crowd workers to crowdsource untranslatability (lexical gaps) and equivalent words between the SL and the TL. Crowd workers, recruited via a chosen crowdsourcing platform, utilize the datasets created earlier.

Given that lexical diversity can occur 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 tuples and their glosses—are mapped to the TL. Here, crowd workers identify equivalent terms and lexical gaps in the TL. For example, Figure 7.2 (a) illustrates an experiment conducted from SL English language to the TL Arabic language, as described in our English-Arabic experiments in Section 9.1. The second experiment

reverses this direction, considering the TL from the first experiment as the SL, and vice versa. In this phase, previously identified equivalents (overlapping terms) are excluded, and the focus shifts to mapping the remaining TL entries to the SL. Figure 7.2 (b) provides an example of this second experiment, conducted from Arabic SL to English TL.



(a) Matching of SL English words with TL Arabic words.

(b) Matching of SL Arabic words with TL English words.

Figure 7.2: Crowdsourcing experiments between SL and TL in two directions.

Task crowdsourcing involves three steps: (1) crowd selection, (2) contribution collection, and (3) contribution quality control. These steps are described as follows:

Crowd selection.

Selecting high-quality crowd workers is vital for ensuring the accuracy of their responses. To achieve this, we suggest the following procedures[1].
1. *Proficiency Test:* The proficiency test assesses the abilities of individuals in crowdsourced tasks, usually comprising 10% to 25% of the total questions of a crowdsourcing task [99]. It should be conducted on the same platform as the actual tasks.

[1]On platforms like Prolific or MTurk, requesters can leverage built-in features [128] for controlling the quality of their crowd.

2. *Crowd Filtering using Alpha:* The requester employs Krippendorff's Alpha (detailed in Appendix A) to identify low-quality workers through inter-annotator agreement (IAA), due to its ability to handle chance agreement, account for multiple annotators, manage missing data, and provide a nuanced measure of both agreement and disagreement. Proficient workers, along with a linguistic expert, answer 25% of the total questions related to SF source words. Then, our method for crowd filtering using Alpha is applied. A flowchart of the method is depicted in Figure 7.3.

The process for filtering workers who have successfully passed the proficiency test involves the collaboration of a linguistic expert and the use of LingoGap, a crowdsourcing platform developed specifically for collecting diversity data (further details are provided in the next section). The procedure begins with the identification of SL, TL, and SF, as well as the selection of a short list of SL terms that belong to the SF, typically comprising around 10 words. Additionally, the IAA threshold is determined, along with the compilation of the list of qualified workers and the linguistic expert. Subsequently, the requester employs LingoGap to generate a crowdsourcing task that incorporates the identified SL terms.

Next, the workers perform the task. If Alpha for their responses is greater than or equal to the threshold, the requester registers these workers as high-quality workers. However, if Alpha is less than the threshold—which represents the worst case—a subset of workers, along with the expert, is required to perform the task. This subset is generated by taking all possible combinations of workers, ranging from the entire list minus one worker to a single worker, utilizing the mathematical combination function [32]. For instance, comb(i) of a set of n workers yields all possible subsets of size i from the list. For each combination, beginning with subsets of size (n-1), the expert and the selected workers conduct the task. If Alpha for their responses is greater than or equal to the threshold, the requester categorizes the subset of workers as high-quality, while the remaining workers (the original list minus the selected subset) are classified as low-quality.

For illustration, consider a group of workers (G_1) comprising three individuals $\{w_1, w_2, w_3\}$ who have passed the proficiency test, along with Exp, a linguistic expert. The process begins with the three workers who perform the task and the requester calculates the Alpha value for their responses. If this value exceeds the predetermined threshold, the requester designates the workers as high-quality. However, if the Alpha value falls below the threshold, the process initiates a step to identify the low-quality worker. In this scenario, the task is repeated with Exp and two members of G_1. Worker subsets of size 2, generated using the combination function [32], include $\{w_1, w_2\}$, $\{w_1, w_3\}$, and $\{w_2, w_3\}$. The expert, along with each of these subsets—$\{Exp, w_1, w_2\}$, $\{Exp, w_1, w_3\}$, and $\{Exp, w_2, w_3\}$—performs the task.

Consider, for instance, the combination $\{Exp, w_1, w_2\}$. If the Alpha

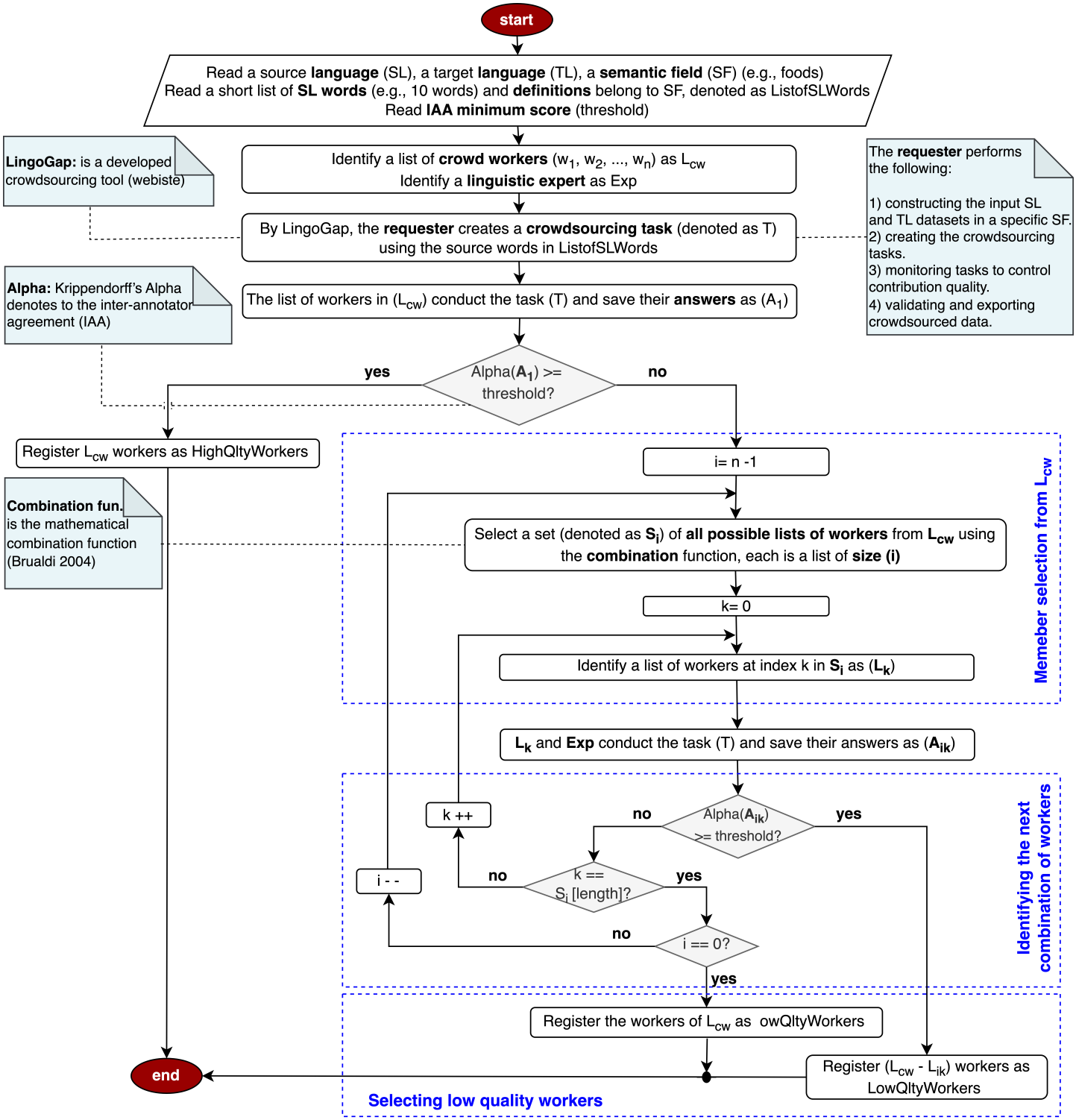


Figure 7.3: Crowd Filtering using Alpha.

value for this group meets the threshold, w_3 is identified as low-quality, and the process terminates. If the Alpha value for all three combinations is below the threshold, the task is repeated with the expert and a single member of G_1. The combinations of size 1 include $\{w_1\}$, $\{w_2\}$, and $\{w_3\}$. The expert, in conjunction with each of these combinations—$\{Exp, w_1\}$, $\{Exp, w_2\}$, and $\{Exp, w_3\}$—performs the task. For example, if the combination $\{Exp, w_1\}$ yields an Alpha value that meets the threshold, w_2 and w_3 are flagged as low-quality, and the process terminates. If all combinations across the three sizes indicate low quality, all workers in G_1 are classified accordingly, thereby concluding the crowd filtering process.

Contribution collection.

The process of identifying gaps and equivalent words between the SL and the TL is conducted using a crowdsourcing platform called *LingoGap*, which is a web-based tool accessible through a server link or commercial platforms like Prolific. It offers two primary roles: admin (e.g., researcher) and worker. The admin is responsible for importing datasets, managing source and target word lists, and creating tasks through a dedicated interface. Clear guidelines for workers are a key focus, supported by a flexible spreadsheet template that allows customization of instructions. Tasks are structured as multi-step queries based on the chain-of-thought (CoT) methodology [96], which breaks down complex linguistic tasks into manageable steps. This approach enables non-expert contributors to complete tasks efficiently, enhancing the reliability of the crowdsourcing process.

Workers interact with LingoGap's interface to respond to multi-step multiple-choice questions (MCQs) for each word. These questions assess whether the SL meaning exists in the TL, identify equivalent meanings from a provided list, or allow workers to add missing meanings. This MCQ format has proven effective for lexical semantic tasks, as demonstrated by prior research [157]. The structured task design enables workers to systematically record lexical gaps or identify equivalent words, ensuring both accuracy and consistency in the collected data. The LingoGap system, along with the roles and tasks of the admin and workers, is described in more detail in Section 8.2.

Contribution quality control.

Implementing live quality control mechanisms is crucial for ensuring the reliability of crowdsourced data and the success of crowdsourcing projects. We use the following tools, illustrated in Figure 7.4, to achieve this goal:

Attention Check Question (ACQ). We use ACQs to gauge worker attentiveness and reliability. These questions are disguised as regular tasks but

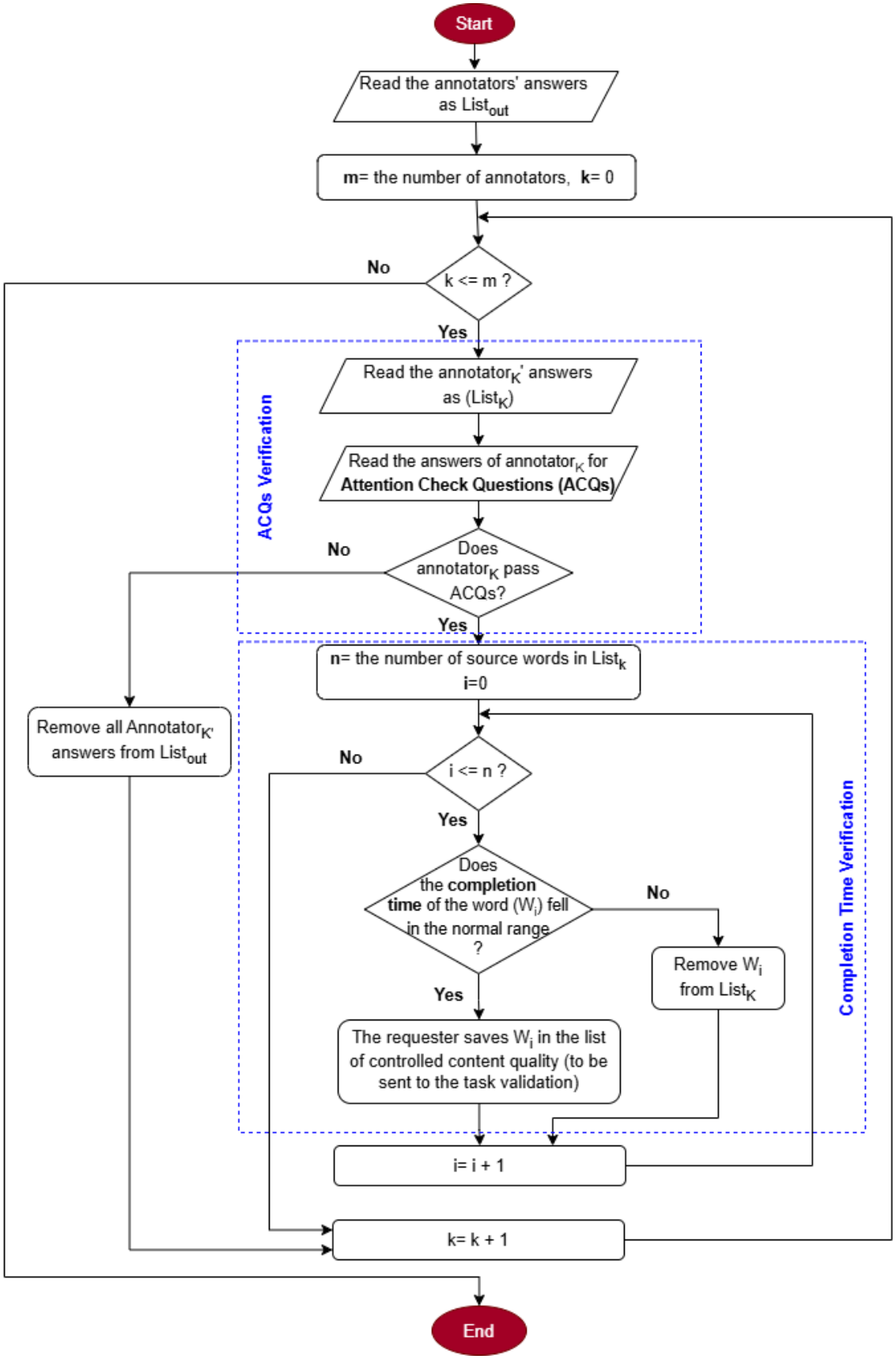


Figure 7.4: Overview of output quality control method.

serve to ensure contributors carefully follow instructions and provide accurate responses. ACQs typically include straightforward queries related to task instructions or preceding content, designed to filter out workers who may rush through tasks. In our approach, we follow the criterion set by Liu et al. [99], presenting one ACQ for every 10 questions, and workers must answer at least 90% of ACQs correctly, aligning with findings by Robinson et al. [128].

Completion Time is the duration taken by a contributor to complete a question in a crowdsourced task, serving as a measure of task efficiency. Monitoring completion time helps evaluate contributor performance, with unusually short times suggesting inattentiveness or rushed responses. LingoGap records response details, including completion time, in its database, excluding outliers significantly deviating from the worker's average completion time. For instance, responses with completion times substantially shorter or longer than the worker's average are excluded from answers.

7.3 Step 3: task validation

In this step, we validate crowdsourced gaps and words using two processes: automatic validation, employing Krippendorff's Alpha to gauge inter-annotator agreement (IAA) among workers' answers and filtering out those with low IAA, and expert validation, where a linguistic expert reviews responses with low IAA.

Automatic validation.

IAA mechanisms are widely utilized to evaluate the reliability of crowdsourcing tasks within computational linguistics [10]. Notably, statistical measures such as Cohen's Kappa [156] and Krippendorff's Alpha [89] serve as essential tools for validating crowdsourced data, particularly in contexts involving multiple annotators. These metrics assess the degree of consensus among annotators, offering an objective measure of annotation reliability. Krippendorff's Alpha, in particular, is a versatile reliability coefficient applicable across a range of data types, including nominal, ordinal, interval, and ratio scales. In contrast, Cohen's Kappa is specifically designed for assessing agreement between two annotators on nominal data [124].

In our crowdsourcing methodology, where the number of crowd workers ranges from two to several (e.g., four participants per task), we propose a validation method based on Alpha to measure IAA among multiple workers. An example of the step-by-step calculation of Alpha is detailed in Appendix A. This approach allows the requester to systematically filter out workers whose contributions result in low IAA, ensuring that only answers with less than 100% IAA from high-quality data are forwarded for

expert validation. A visual representation of this method is provided in the flowchart shown in Figure 7.5.

The process for validating data crowdsourced from a group of workers (G_1) involves the collaboration of a second group of crowd workers (G_2) and the use of the LingoGap platform. The procedure begins by reading SL, TL, and SF, along with a list of answers corresponding to a specific task, which includes SL terms and their respective definitions. The IAA threshold is then identified, with the workers in G_1 denoted as $\{w_1, w_2, \ldots, w_n\}$, and the workers in G_2 as $\{w\prime_1, w\prime_2, \ldots, w\prime_m\}$. A linguistic expert, referred to as *Exp*, is also identified for the validation process.

Subsequently, the requester calculates Alpha using a custom-developed Python script. If the Alpha value for the responses from G_1 workers exceeds the predetermined threshold, the workers are classified as high-quality, and their responses with less than 100% IAA are forwarded for expert validation. Conversely, if the Alpha value falls below the threshold—indicating the worst-case scenario—the requester employs LingoGap to generate a crowdsourcing task incorporating the specified SL terms and their definitions. A subset of workers from G_1, as well as a subset from G_2, is required to perform this task. The subset from G_1 is generated by computing all possible combinations of workers, ranging from (n-1) workers, where n is the total number of G_1 workers, down to a single worker, using the mathematical combination function [32]. The same method is applied to G_2, generating subsets from a single worker up to m workers, where m is the total number of G_2 workers. For each subset of G_1 workers, beginning with size (n-1) and decreasing to 1, and for each subset of G_2 workers, beginning with size 1 and increasing to m, the workers perform the task. If the Alpha value for the responses surpasses the threshold, the requester classifies the selected subset of G_1 workers as high-quality, while the remaining G_1 workers—those not included in the subset—are categorized as low-quality.

Consider, for example, a group of workers (G_1) comprising three individuals, $\{w_1, w_2, w_3\}$, and another group (G_2) consisting of three individuals, $\{w\prime_1, w\prime_2, w\prime_3\}$, along with *Exp*, a linguistic expert. The validation process begins with the requester calculating the Alpha value for the responses provided by $\{w_1, w_2, w_3\}$. If the calculated Alpha exceeds the predefined threshold, the workers $\{w_1, w_2, w_3\}$ are classified as high-quality, and their responses with less than 100% IAA are submitted to the expert for further validation. However, if the Alpha value falls below the threshold, the process moves to the identification of low-quality workers. At this stage, the task is repeated with two workers from G_1 and one worker from G_2. Subsets of size 2 from G_1 are generated using the combination function, resulting in the following groups: $\{w_1, w_2\}$, $\{w_1, w_3\}$, and $\{w_2, w_3\}$. Similarly, subsets of size 1 from G_2 are generated, yielding $\{w\prime_1\}$, $\{w\prime_2\}$, and $\{w\prime_3\}$. The resulting combinations of workers from G_1 and G_2 are $\{w_1, w_2, w\prime_1\}$, $\{w_1, w_3, w\prime_1\}$, $\{w_2, w_3, w\prime_1\}$, $\{w_1, w_2, w\prime_2\}$, $\{w_1, w_3, w\prime_2\}$, $\{w_2, w_3, w\prime_2\}$, $\{w_1, w_2,$

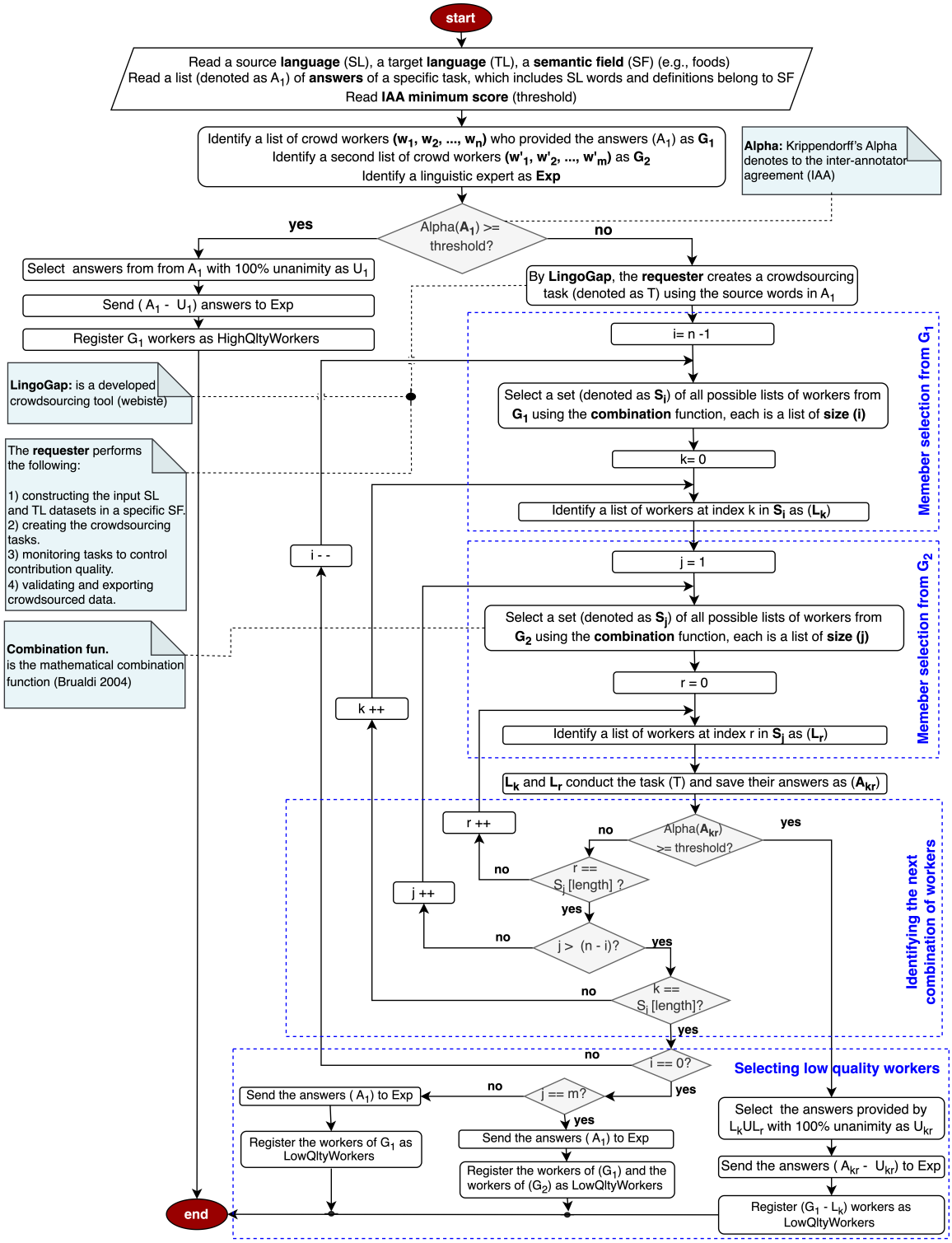


Figure 7.5: Crowdsourced Data Validation using Alpha.

$w\prime_3\}$, $\{w_1, w_3, w\prime_3\}$, and $\{w_2, w_3, w\prime_3\}$. Each subset, beginning with $\{w_1, w_2, w\prime_1\}$ and proceeding through to $\{w_2, w_3, w\prime_3\}$, performs the task if the Alpha value of the preceding combination in the sequence fails to meet the threshold.

For instance, the subset $\{w_1, w_2, w\prime_1\}$. If the Alpha value for this subset meets the threshold, w_3 is identified as a low-quality worker, and their responses with less than 100% IAA are forwarded to *Exp* for validation, concluding the process. Conversely, if the Alpha values for all nine subsets fall below the threshold, the task is repeated using a single worker from G_1 and two workers from G_2. The combinations of size 1 from G_1 are $\{w_1\}$, $\{w_1\}$, and $\{w_3\}$, while the combinations of size 2 from G_2 include $\{w\prime_1, w\prime_2\}$, $\{w\prime_1, w\prime_3\}$, and $\{w\prime_2, w\prime_3\}$. The resulting combinations between G_1 and G_2 in this stage are $\{w_1, w\prime_1, w\prime_2\}$, $\{w_2, w\prime_1, w\prime_2\}$, $\{w_3, w\prime_1, w\prime_2\}$, $\{w_1, w\prime_1, w\prime_3\}$, $\{w_2, w\prime_1, w\prime_3\}$, $\{w_3, w\prime_1, w\prime_3\}$, $\{w_1, w\prime_2, w\prime_3\}$, $\{w_2, w\prime_2, w\prime_3\}$, and $\{w_3, w\prime_2, w\prime_3\}$. Beginning with the subset $\{w_1, w\prime_1, w\prime_2\}$ and progressing through to $\{w_1, w\prime_2, w\prime_3\}$, each combination performs the task, provided the Alpha value of the preceding subset remains below the threshold. For example, if the combination $\{w_1, w\prime_1, w\prime_2\}$ produces an Alpha value that surpasses the threshold, both w_2 and w_3 are designated as low-quality, and their responses with less than 100% IAA are sent to *Exp*, thus concluding the process. Finally, if the Alpha values for all nine combinations remain below the threshold, the task is repeated with the workers in G_2 only, specifically $\{w\prime_1, w\prime_2, w\prime_3\}$. If Alpha for this set exceeds the threshold, all workers in G_1 are classified as low-quality, and their answers with less than 100% IAA are sent to *Exp*, thereby terminating the process. If Alpha for this final combination is also below the threshold, both G_1 and G_2 workers are deemed low-quality, and the original list of responses from G_1 is forwarded to *Exp*, thus completing the validation process.

Expert validation.

A bilingual expert checks answers from automatic validation that did not achieve 100% IAA. A spreadsheet is created with columns: Worker-ID, Source Lemma, Source Gloss, and Worker's answer (lexical gap, equivalent word, or 'Do not know'). Also, 10% of words in the spreadsheet are included for the expert's "sanity check", chosen based on 100% IAA. This sheet is provided to a linguistic expert, who is asked to perform tasks as follows: (1) equivalent meanings: mark provided target language words as correct or incorrect. In the case of an incorrect word, the expert provides the correct word or indicates it as a gap. (2) lexical gaps: confirm or reject source language words marked as gaps or non-gaps in the target language by the crowd. In the case of an incorrect gap, the expert provides the appropriate word. and (3) do not know: provide an answer for source words, marking them as gaps or identifying equivalent words in the target language.

8 LingoGap Platform

Contents

This chapter provides an in-depth exploration of LingoGap, a novel crowdsourcing tool developed to identify linguistic gaps and equivalent terms across languages. The discussion begins with an overview of the LingoGap environment in Section 8.1, describing its foundational components and the design principles that support its functionality. Section 8.2 focuses on the LingoGap System, explaining its operational mechanics and the integration of user contributions. A critical assessment of the tool's usability is presented in Section 8.3, highlighting how user-centric design enhances interaction efficiency through UX/UI evaluation. Finally, Section 8.4 explores the application of LingoGap within the DataScientia platform, showcasing its practical utility in addressing multilingual challenges by creating multilingual, diversity-aware datasets for real-world scenarios.

8.1 LingoGap environment

This section describes the working environment of LingoGap, where the key components are illustrated in Figure 8.1. The LingoGap website serves as the central hub for the crowdsourcing process, interacting with a central *database* to read and write data. This database functions as a repository for storing input datasets from language pairs (source and target languages) within a semantic domain, as well as recording task results—identified equivalent terms and lexical gaps.

LingoGap facilitates collaborative efforts by leveraging contributions from two distinct user roles: *task requesters* and *crowd workers*. Task requesters are responsible for creating and designing crowdsourcing tasks. They oversee

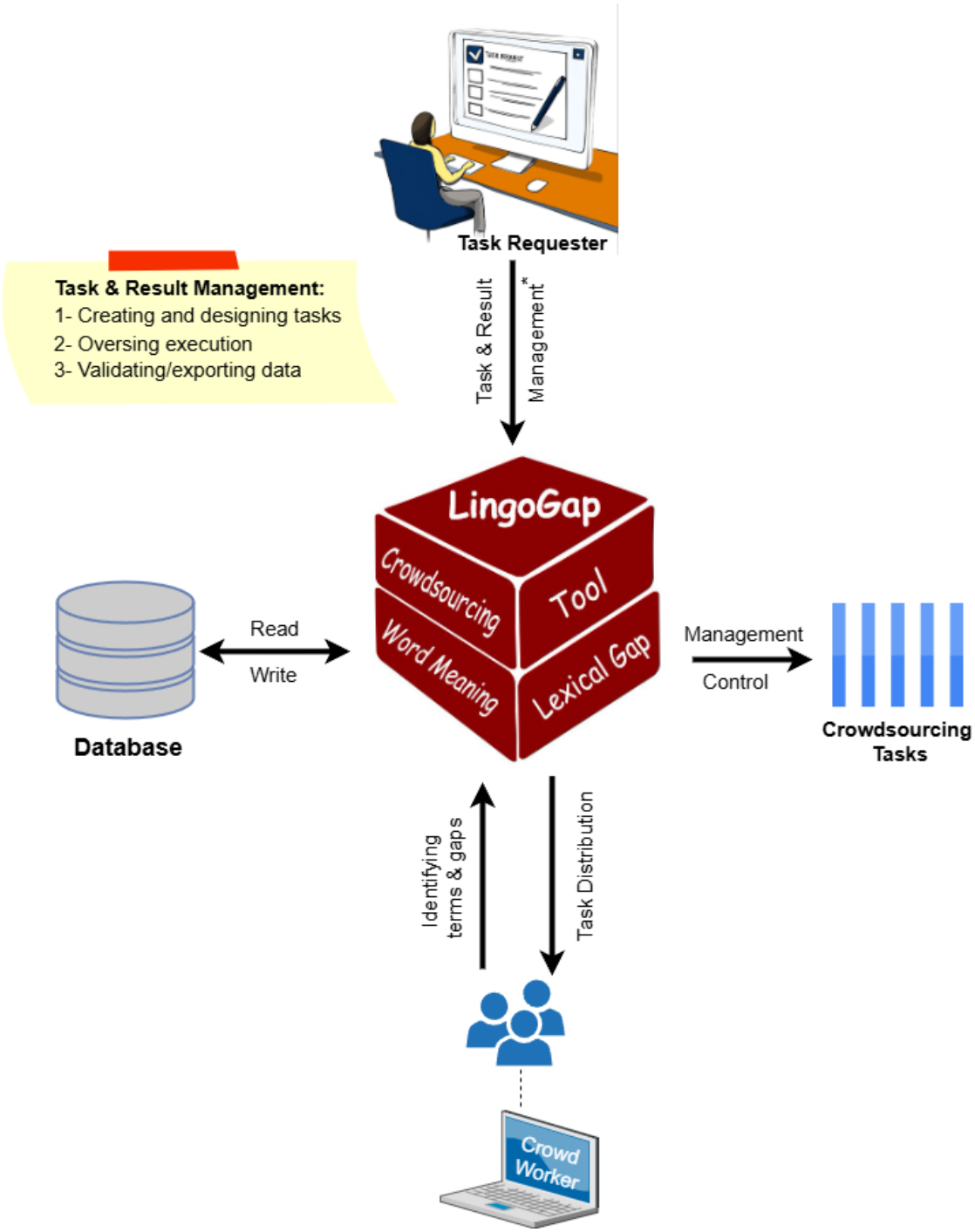


Figure 8.1: LingoGap Environment

task execution, ensure the quality and consistency of contributions, and validate or export the resulting data. Crowd workers are essential participants who complete linguistic tasks. Their responsibilities include identifying lexical gaps, proposing equivalent word meanings, and providing insights on terms and phrases.

Furthermore, LingoGap supports efficient task distribution and management control. Tasks are systematically assigned to crowd workers, who perform lexical comparisons and contribute their findings for each micro-task. A crowdsourcing *micro-task* typically consists of one lexical entry from the SL and all corresponding lexical entries (each represented as a tuple of a word and gloss) from the TL as candidates.

8.2 LingoGap system

Chosen workers compare SF words in the SL and the TL using the LingoGap crowdsourcing tool to identify gaps and equivalent words in both languages. LingoGap, developed with Java (JSP), JavaScript, and MySQL, is a crowdsourcing website accessible to workers via a server link. It can also integrate with commercial platforms like Prolific.

LingoGap involves two roles: admin (or requester) and worker. Admin, e.g., a researcher, imports SF datasets into the database and manages tasks. The admin interface[1] includes tabs for '*Experiment*', '*Source Words*', and '*Target Words*'. In the '*Experiment*' tab, the requester sets task details like description, languages, and date, and selects source lemmas along with their glosses from the SF dataset. Source and target word lists are managed in the '*Source Words*' and '*Target Words*' tabs. Tasks can be closed via a dropdown list of active tasks. Providing detailed instructions and clear guidelines for workers is essential. Therefore, we created a spreadsheet template that can be used by the requester during task creation to define guidelines. This template includes two columns: the first contains a list of tips or questions (e.g., are there any restrictions for this experiment?). The second column contains answers to the tips and questions outlined in the first column (e.g., the only restriction is that machine translation is not allowed for defining words). We designed it to be as flexible as possible to streamline content adjustments. For example, it includes nine default guidelines, but the requester can add, edit, or delete records based on the ongoing experiment. Once this spreadsheet is prepared, the requester must import it into the LingoGap database.

In the task design, we employed the multi-step answering approach outlined by Mohammad et al. [107], in which SL-TL comparison tasks were decomposed into smaller and simpler queries. Workers were tasked with evaluating whether a given SL word is associated with an equivalent TL

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

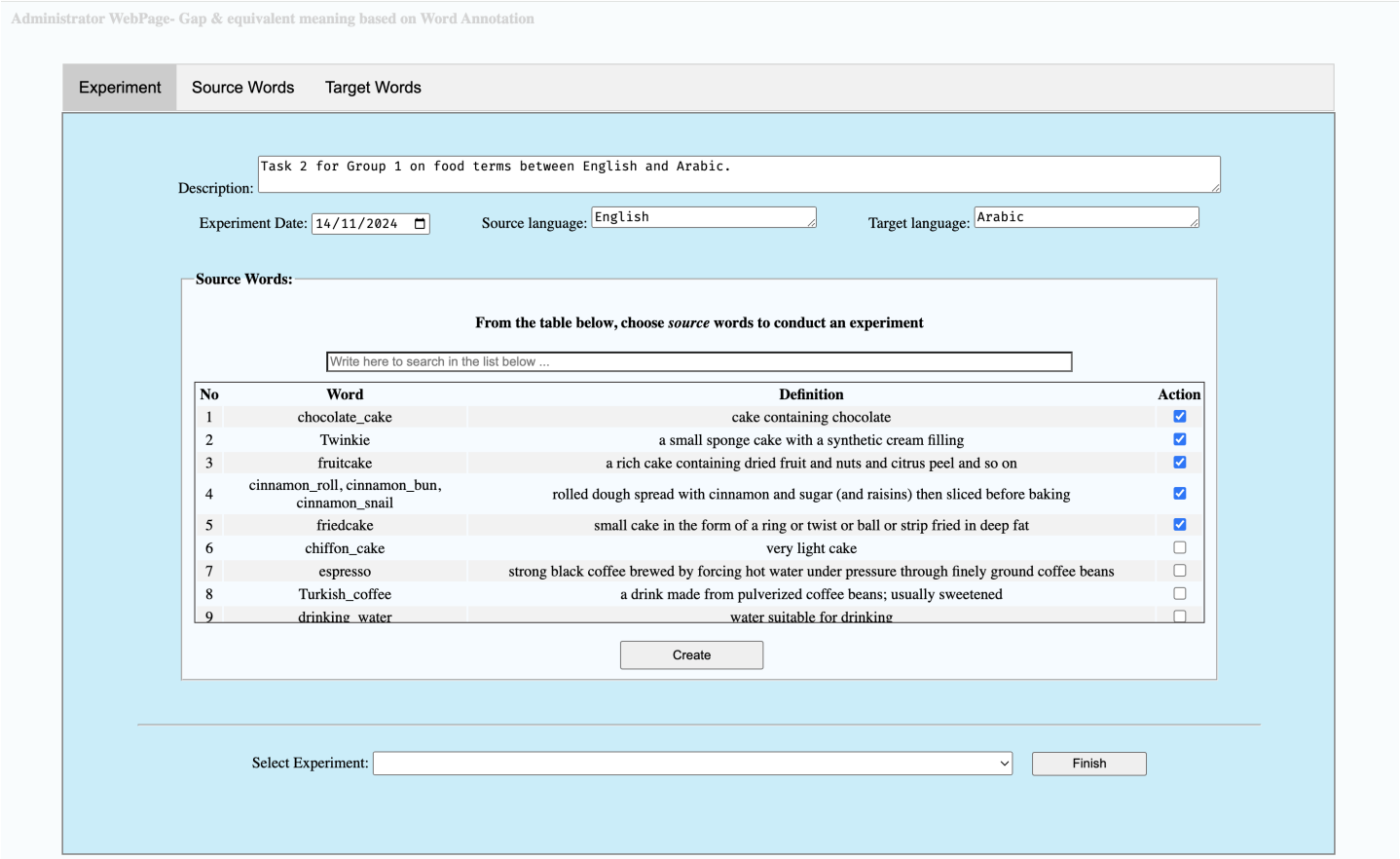


Figure 8.2: Admin's GUI

word. This structured approach, also known as the chain-of-thought (CoT) method [96], facilitates the logical progression of tasks, allowing non-expert contributors to complete them more effectively and ultimately enhancing the overall performance of crowdsourcing tasks.

After task creation, crowd workers use LingoGap's worker interface[2] and follow guidelines to answer multi-step questions for each source word presented sequentially. Each word prompts three multiple-choice questions (MCQs), where the utility of MCQs for domain-targeted tasks has been demonstrated by Welbl et al. [157] for conducting lexical semantic tasks. The three choices are: 1) Determine if the target language includes the meaning. If not, this meaning is recorded as a lexical gap in the target language. 2) If included, compare meanings, the source meaning and a list of target meanings, and choose the equivalent from the list. 3) If included in the target language and the meaning is not listed, add the missing target

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

Figure 8.3: Crowdworker's GUI

lemma and gloss. Both interfaces were utilized to conduct our experiments, as detailed in Chapter 9.

8.3 UX/UI evaluation

The UX/UI evaluation of LingoGap aimed to assess user interactions with the platform's interface through a structured survey[3]. This survey collected feedback on key user experience and usability factors, including the clarity of instructions, ease of navigation through multi-step tasks, visual appeal, interface responsiveness, and overall user satisfaction.

8.3.1 Survey design and scope

Participants answered a series of questions addressing distinct stages of the LingoGap task flow. Each question targeted a specific aspect of the user experience to gather insights on (1) the clarity of instructions, (2) the ease

[3] https://docs.google.com/forms/d/e/1FAIpQLSd7fn9uoY0Qzsmy2Wo5O_ZT7g4AhSFbMkfTMxl0k2RCIaecUQ/viewform

of following steps for lexical evaluation, (3) visual and interaction design elements, and (4) the efficiency of task completion within the interface. Rating scales were employed to quantify feedback on aspects such as instruction clarity (ranging from "*Very Clear*" to "*Very Confusing*"), the usability of specific steps, and interface responsiveness.

8.3.2 Key aspects evaluated

The survey addressed the following key aspects:

1. *Instructional clarity*: questions were designed to evaluate how well participants understood the instructions, particularly for multi-step lexical evaluations. This aspect is critical, as unclear instructions can significantly impact task accuracy and user efficiency.

2. *Task process and ease*: participants rated the ease of navigating the multi-step process, which includes comparing source and target meanings, selecting equivalent terms, and adding missing lexical items. Understanding usability at each step helps identify potential barriers or points of confusion..

3. *Visual and interaction design*: the survey assessed the visual appeal of the layout, including font size, color contrast, and overall readability. These factors contribute to user satisfaction and ease of interaction.

4. *Responsiveness and efficiency*: the responsiveness of interactive elements, such as buttons, was evaluated alongside users' perceptions of task completion efficiency. This aspect reflects the interface's ability to support users in completing tasks effectively.

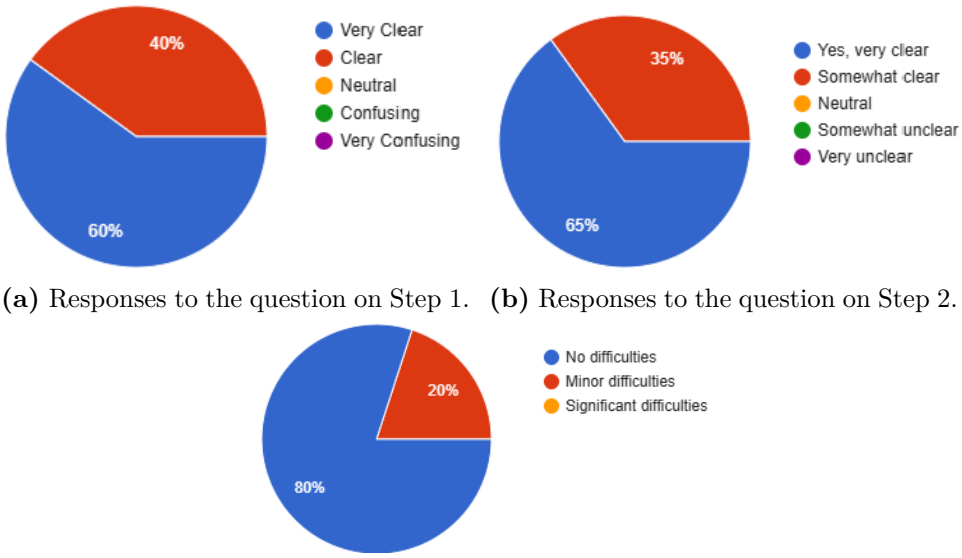
5. *Overall user satisfaction and recommendation*: finally, participants rated their overall experience to recommend LingoGap, providing a comprehensive perspective on user satisfaction.

8.3.3 Findings and implications for UX/UI design

On the positive side, our method demonstrates a robust ability to generate diversity-aware datasets across various semantic fields in two distinct languages. Additionally, it effectively accommodates datasets of varying scales. In larger domains with substantial data, as shown in our experiments (described in Chapter 9), the method enables efficient dataset partitioning. For instance, in the experiments detailed in Section 9.1, 35 SL words were allocated per crowdsourcing task (e.g., 35 words from a pool of 2,364 English words in the English-to-Arabic experiment). Similarly, in the experiments discussed in Section 9.2, 45 SL words were assigned per task (e.g., 45 words from 1,448 Indonesian words in the Indonesian-to-Banjarese experiment).

Another example from the kinship domain, described in Section 9.3, involves the allocation of kinship terms for crowdsourcing tasks. In the English-to-Persian experiment, 17 English kinship terms were allocated per task, while in the reverse direction, 30 Persian kinship terms were allocated per task.

This scalability highlights the versatility of our approach, allowing efficient crowdsourcing to achieve lexical diversity across both large and small semantic domains in our case studies. Additionally, this approach shows potential for broader application to other languages and domains.



(a) Responses to the question on Step 1. **(b)** Responses to the question on Step 2.

(c) Responses to the question on Step 1.

Figure 8.4: Questions on a three-step crowdsourcing process using LingoGap.

The survey responses indicate that the overall process of using LingoGap to compare source words with target words for identifying equivalents and lexical gaps is rated as "*Very Satisfied*" on average. This result is based on three questions related to the three-step diversity data crowdsourcing process, which were included in the survey. The first question pertains to Step 1, asking crowd workers how easy and clear it was to determine whether the target language includes an equivalent meaning. The responses show 60% rated it as "*Very Clear*" and 40% as "*Clear*", as illustrated in Figure 8.4 (a). The second question focuses on Step 2, which evaluates whether the definitions in the list were easy to understand when choosing the equivalent meaning. According to Figure 8.4(b), 65% of respondents rated it as "*Very Clear*" and 35% as "*Clear*". Finally, the third question addresses Step 3, asking crowd workers if they encountered difficulties in selecting or adding missing target lemmas. As shown in Figure 8.4(c), 80% reported "*No Difficulties*", while 20% indicated "*Minor Difficulties*".

For the graphical user interface (GUI), aspects such as visual appeal (e.g., font size and color contrast) and button responsiveness received high ratings. While the interface was generally assessed as "*Very Appealing and Clear*" and "*Very Responsive*", several suggestions were provided by both requesters and workers to further enhance LingoGap:

1. Adding a *word count* in the data table displaying the SL lemmas and glosses on the admin page to streamline task creation by aligning this data with the source spreadsheet.

2. Recording the *date and time*, alongside the *duration of responses*, to facilitate the analysis of contribution quality.

3. Incorporating a '*back*' button into the GUI of selecting target words in the second step of answering, as well as into the GUI of proposing new words in the third step. This would allow workers to return to the initial step if they mistakenly answered the first question.

4. Including multiple gap *examples* instead of just one and providing clearer restrictions in the guidelines, such as explicitly prohibiting the use of machine translation tools like Google Translate.

Based on pilot and initial feedback, we improved LingoGap: an admin word count feature, worker '*back*' button (see Figure 8.3), and answer date recording were added. We also enhanced guideline flexibility for requesters, allowing them to freely manage instructions and gap examples.

Furthermore, we encountered certain limitations during our experiments described in Chapter 9. For instance, the timing for task completion is critical. In response to feedback from the crowd, we temporarily suspended crowdsourcing tasks during exam periods and holidays, such as Christmas and Iftar days, which followed five days after Ramadan. Another limitation was the expectation that each group member completed a task within three days. Any delay by a member affected the overall task schedule. For example, Task 10 by G_4 in the Arabic-to-English experiment was delayed by two weeks because w'_3 was on a fourteen-day rest period after giving birth. Another example, in the Indonesian-to-Banjarese experiment, Task 7 by G_1 was halted for two days because one of the participants had to attend their family event. Our experience confirmed previous findings that the time during which one executes a crowdsourcing campaign is critical, as real world situations can affect workers' responses in a number of ways [39].

8.4 Using LingoGap via DataScientia

DataScientia[4] is an innovative research initiative hosted by the Department of Information Engineering and Computer Science (DISI) at the University

[4] https://datascientia.disi.unitn.it/

of Trento. It serves as a multidisciplinary hub, bringing together expertise from computational linguistics, data science, and artificial intelligence to address complex societal challenges. With a focus on advanced data-driven methodologies, DataScientia fosters collaboration among academia, industry, and civil society to design and implement impactful digital solutions. The initiative prioritizes openness, reproducibility, and inclusivity, aiming to develop tools and frameworks that expand access to knowledge while advancing the frontiers of technology and human-centered research.

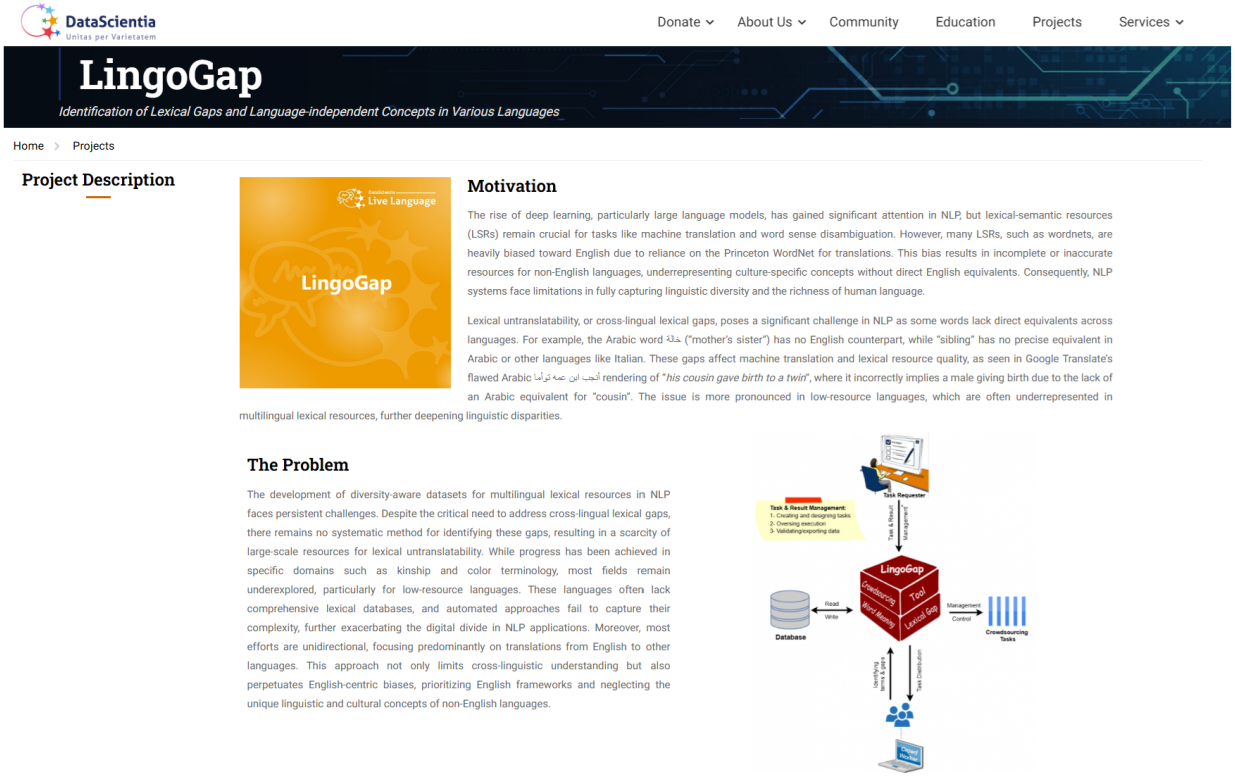


Figure 8.5: The homepage of the LingoGap project on DataScientia

DataScientia offers an integrated ecosystem of services designed to support interdisciplinary research and innovation. Its core components include *LiveMedia* for multimedia data analysis, *LiveLanguage* for linguistic resource development, *LiveKnowledge* for knowledge management and collaboration, *LiveData* for scalable big data analysis, and *LivePeople* for fostering human-centric, participatory approaches. Together, these services integrate technology, data, and human expertise to effectively address complex research challenges.

DataScientia encompasses over sixty projects, engaging thousands of participants from diverse backgrounds. Within the LiveLanguage area, the

LingoGap project[5] provides a comprehensive overview, including detailed project descriptions, case studies, experimental insights, team information, and other relevant details. Figure 8.5 displays the homepage of the LingoGap website.

The LingoGap website provides researchers with essential resources for accessing the LingoGap system to conduct crowdsourcing experiments. Specifically, it includes a link to the requester's GUI for task creation and management, as well as a link to the crowdworker's GUI for task participation. Additionally, the website offers documentation, including a step-by-step user manual for effectively using LingoGap. Figure 8.6 provides a screenshot illustrating this information.

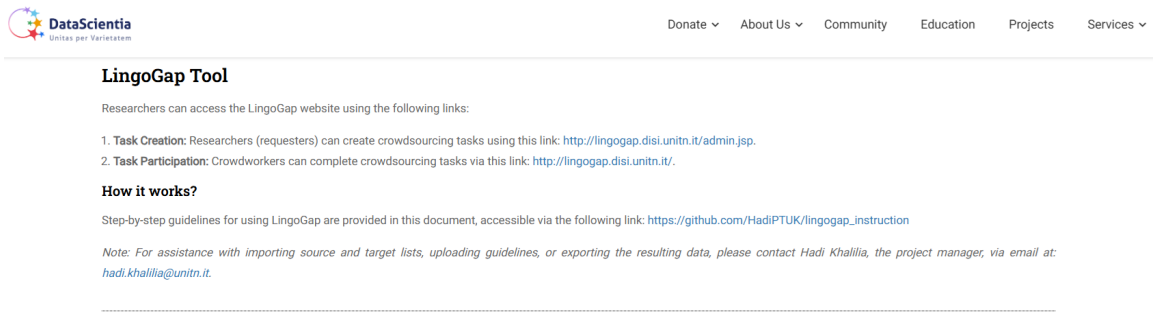


Figure 8.6: A screenshot illustrating the links on DataScientia for using LingoGap

The expert-sourcing lexical diversity project[6] is another example within the LiveLanguage domain, showcasing the expert-driven methodology introduced in Chapter 4. It highlights the case studies and experiments conducted using this methodology, as detailed in Chapter 5. Furthermore, Figure 8.7 displays the homepage of the expert-sourcing lexical diversity website.

[5] https://datascientia.disi.unitn.it/projects/lingogap/
[6] https://datascientia.disi.unitn.it/projects/expert-sourcing-lexical-diversity/

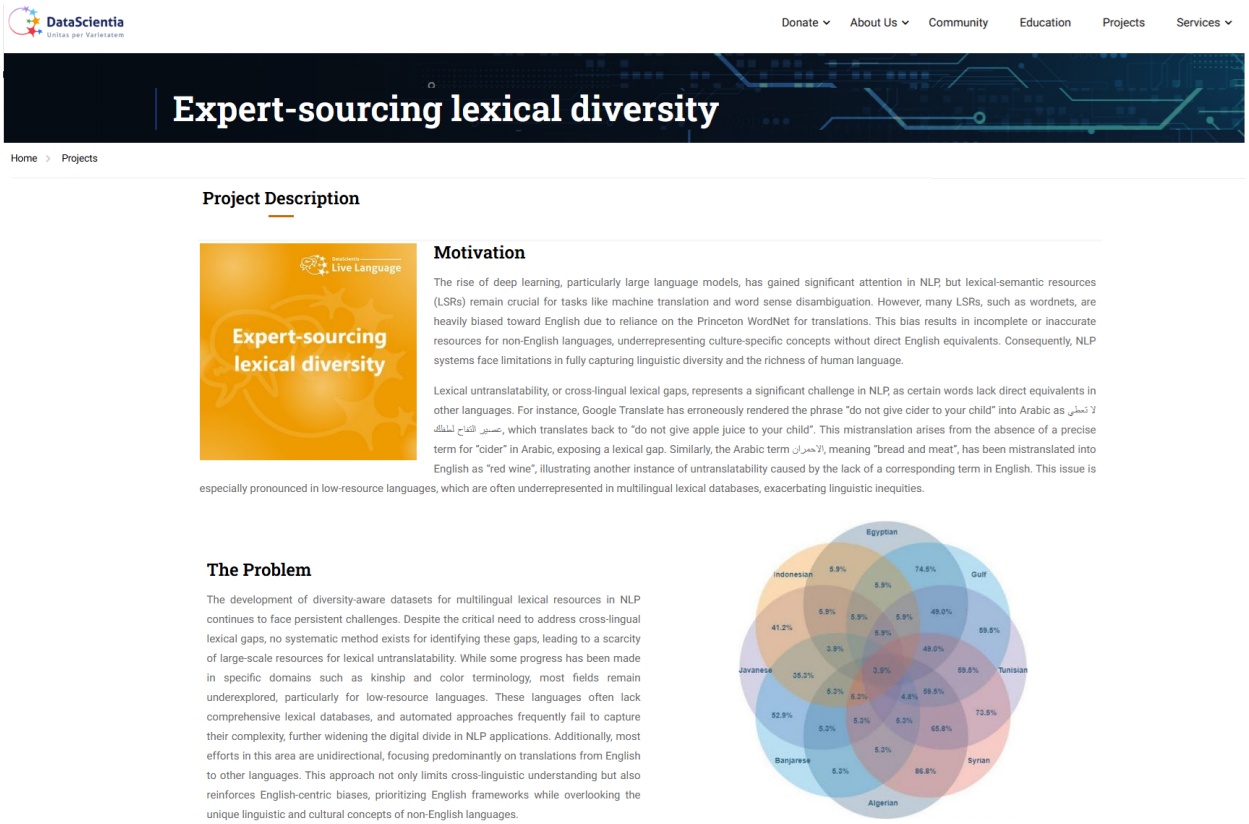


Figure 8.7: The homepage of the expert-sourcing lexical diversity project on DataScientia

9

Crowdsourcing Experiments

Contents

This chapter presents a comprehensive validation of our crowdsourcing methodology through nine experiments, each designed to investigate linguistic and cultural diversity. These experiments provided a structured framework to evaluate the effectiveness and reliability of our approach across multiple languages and semantic domains. Each experiment was supervised by a requester—one of our collaborators, a native speaker of the targeted language—who managed the crowdsourcing environment, ensured consistent participant engagement, and guaranteed the quality of the crowdsourced contributions. The requester played a crucial role in designing task workflows, monitoring progress, and maintaining the integrity of the crowdsourcing process. The experiments are grouped into two primary semantic domains—food and kinship—both deeply tied to culture and language. The experiments are organized as follows:

- *Experiment 1 (English vs. Arabic food)*: this experiment explored cross-cultural and linguistic representations of food-related concepts between English and Arabic speakers.
- *Experiment 2 (Indonesian vs. Banjarese food)*: this study examined food-related terminologies and interpretations within a regional context, focusing on Indonesian and Banjarese.
- *Experiment 3 (English vs. Persian kinship)*: shifting to the kinship

domain, this experiment analyzed the linguistic and conceptual similarities and differences between English and Persian kinship terms.

- *Experiment 4 (Arabic vs. Persian kinship)*: this experiment compared kinship-related linguistic constructs between Arabic and Persian speakers.
- *Experiment 5 (English vs. Turkish kinship)*: this study evaluated how kinship concepts are expressed across English and Turkish.
- *Experiment 6 (kinship concepts vs. Turkish)*: moving beyond direct language comparisons, this experiment examined the alignment of universal kinship concepts with the Turkish language.
- *Experiment 7 (kinship concepts vs. Persian)*: similarly, this study investigated the relationship between universal kinship concepts and their representation in Persian.
- *Experiment 8 (kinship concepts vs. Maba)*: through direct language comparisons, this study examined the alignment of universal kinship concepts with the Maba language.
- *Experiment 9 (LLMs as annotators)*: this section investigates the use of Large Language Models (LLMs), such as GPT-4, DeepSeek, and Gemini, as annotation tools for creating diversity-aware datasets. It also compares the accuracy of these datasets with the accuracy of crowdsourced data generated using our method.

Each section in this chapter illustrates the application of our methodology to crowdsource words and gaps through three key steps: task generation, contribution collection and quality control, and task validation. The insights gained from these experiments not only validate our methodology but also enhance our understanding of linguistic diversity, particularly by exploring the significant meaning divergences of food- and kinship-related terms across languages. This chapter highlights the potential of crowdsourcing as a scalable tool for creating diversity-aware datasets.

9.1 Experiment 1 (English vs Arabic: food)

9.1.1 English → Arabic

Task generation

We conducted this experiment in the food domain, crowdsourcing from English to Arabic. Two input datasets were created: the first was the English dataset, which included 2,364 terms, and the second was the Arabic dataset,

containing 1,607 terms. The English terms were collected from the UKC using the Exporter tool, a built-in UKC management service designed to extract data from the UKC. As illustrated in Figure 9.1, the tool was used to gather English words belonging to a specific semantic field (e.g., food terms in this experiment) from the UKC. This tool takes one or more concepts (e.g., food, nutrient) that represent a given semantic field (e.g., food) and the UKC resource as inputs. It then retrieves all lexicalizations associated with those concepts from the respective language lexicon in the UKC.

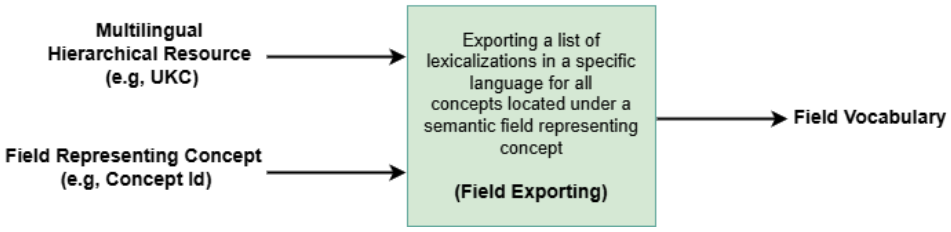


Figure 9.1: Semantic field retrieving method for English.

To address the scarcity of high-quality lexical resources for Arabic, particularly for food-related terms (e.g., UKC, WordNet), we employed a semantic filtering method (depicted in Figure 9.2) based on AraBERT to extract relevant terms from an Arabic online dictionary. As illustrated in Figure 9.3, our method involves three steps: In *Step 1*, a (digital) dictionary and food definitions, including example terms, are inputted into the method, and the inputs are preprocessed by segmenting the paragraphs into sentences. In *Step 2*, the sentences are transformed into vectors using AraBERT. In *Step 3*, dictionary vectors similar to the centroid food vector are clustered using cosine similarity with a threshold of 80%, then re-transformed into text sentences and reported in a spreadsheet. For clarity and ease of understanding, *English* examples were used for the inputs and outputs in the methodology steps instead of Arabic examples. A custom Python script based on AraBERT was implemented following the methodology to collect food-related words in Arabic from Arabic dictionaries (e.g., the Almaany dictionary). The script has been uploaded to GitHub[1].

Crowd selection

(1) A proficiency test: fourteen students, Arabic native speakers and proficient in English, conducted a proficiency test. Five students held bachelor's degrees, while the others had master's degrees in fields like mathematics, computer science, economics, and linguistics. Using LingoGap, they answered 10 food-related questions. Twelve of them, having passed the exam,

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

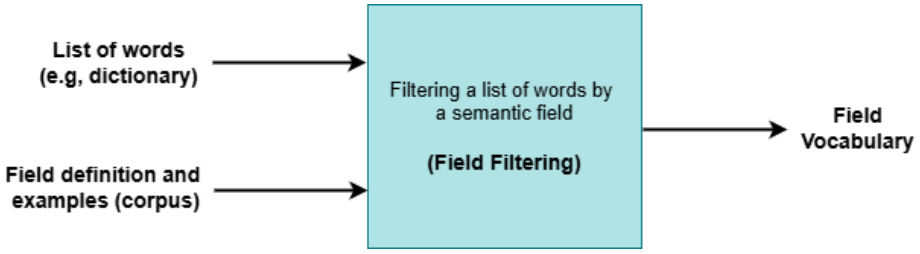


Figure 9.2: Semantic field retrieving method for Arabic.

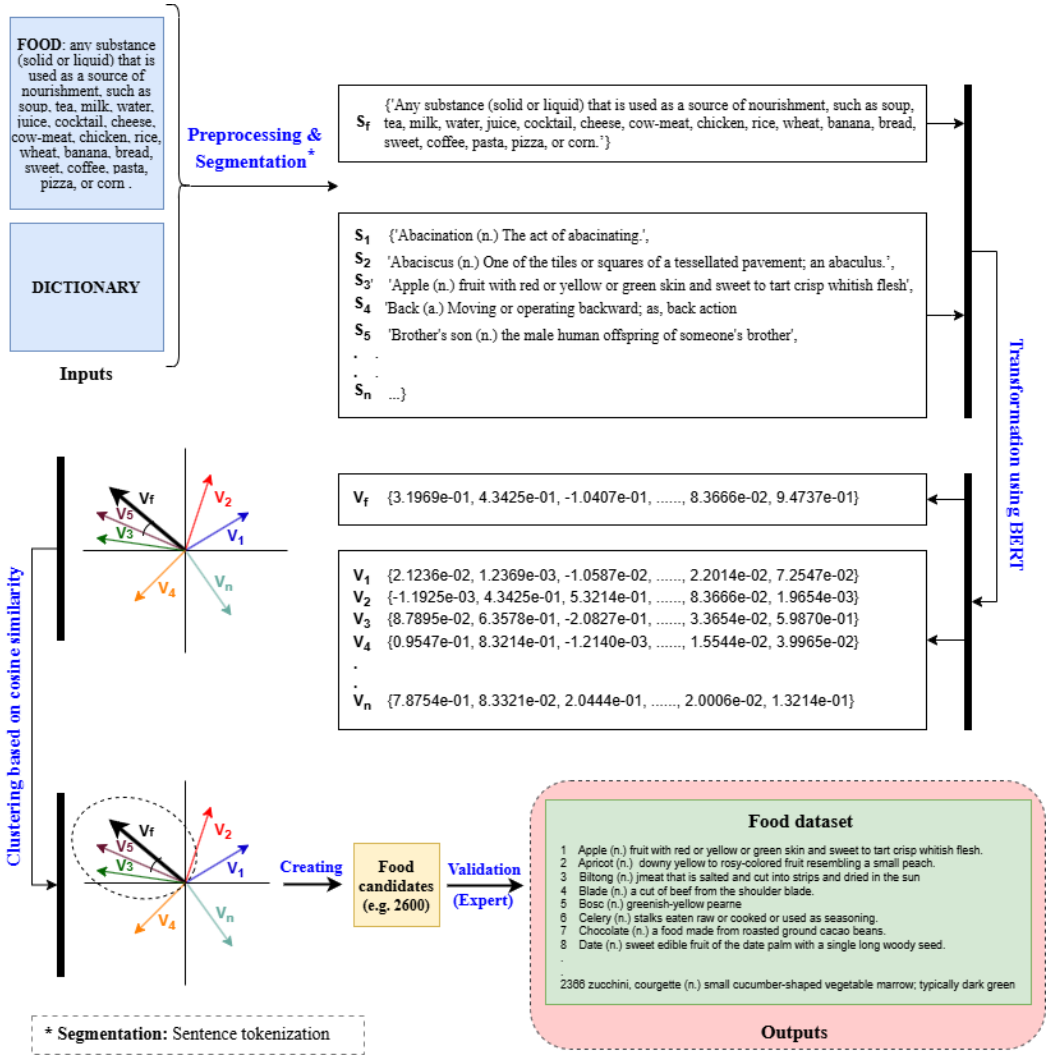


Figure 9.3: Flowchart of the semantic-field filtering method to collect food words from a digital dictionary.

volunteered as crowd workers.

(2) Crowd filtering using Alpha: workers who passed the proficiency exam were split into four groups—G_1, G_2, G_3, and G_4—each consisting of three members. Our methodology for crowd filtering using Alpha, depicted in Figure 7.3, was employed to eliminate low-quality workers not explored during the initial test. Four tasks, each comprising 10 food-related questions, were conducted to evaluate worker quality. G_1 conducted 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.

Contribution collection and quality control

We used the English food dataset, containing 2,364 words, in our experiment. Each participant group, G_1 and G_2, compared 245 English food words with the complete Arabic food dataset across seven crowdsourcing tasks. Similarly, each group, G_3 and G_4, compared approximately 945 English food words (G_3 tasked with 945 words and G_4 with 929 words) with the complete Arabic food dataset across twenty-seven tasks. G_1 and G_3 focused on *alcoholic drinks, pizza, salads, dairy products, rice, bread,* and *fruits*, while G_2 and G_4 worked on *soups, vegetables, cakes, meats, sandwiches,* and *desserts*. *Pilot tasks* were conducted to adjust crowdsourcing parameters and assess LingoGap's usability. As shown in Table 9.1, four tasks with different parameters, such as question count, ACQs, and duration, were performed. Task 3, which had 35 questions, 3 ACQs, and a 60-minute duration, was deemed the most suitable.

Task	Questions	ACQs	Time(min)
1	15	1	25
2	25	2	50
3	35	3	60
4	45	4	90

Table 9.1: Questions count and the time of each pilot task.

Using LingoGap's admin GUI, we created a task (e.g., Task 1 for G_1), specifying English as the source language, Arabic as the target, and selecting 35 English words. Additionally, guidelines were imported at this stage, as detailed in Table 9.2.

Crowd workers logged into LingoGap with their names and then reviewed and applied the guidelines. For each English word, LingoGap presents three steps. *Step 1* shows the English word and asks if Arabic has an equivalent. If confirmed, LingoGap proceeds; if not, it records a gap. *Step 2* displays a list of Arabic words for food word selection, as shown for the example of

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 there is a lexical gap in Arabic.
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. (Optional) 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 70 minutes. 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.

Table 9.2: Guidelines for the English vs. Arabic experiment.

exploring '*banana*' in Figure 9.4. If the Arabic word is not found in the list, *Step 3* allows the crowd to input an equivalent meaning into text boxes.

Three ACQs were utilized for real-time quality control of crowdsourced gaps and words in each task. For example, in Task 6 conducted by G_2,

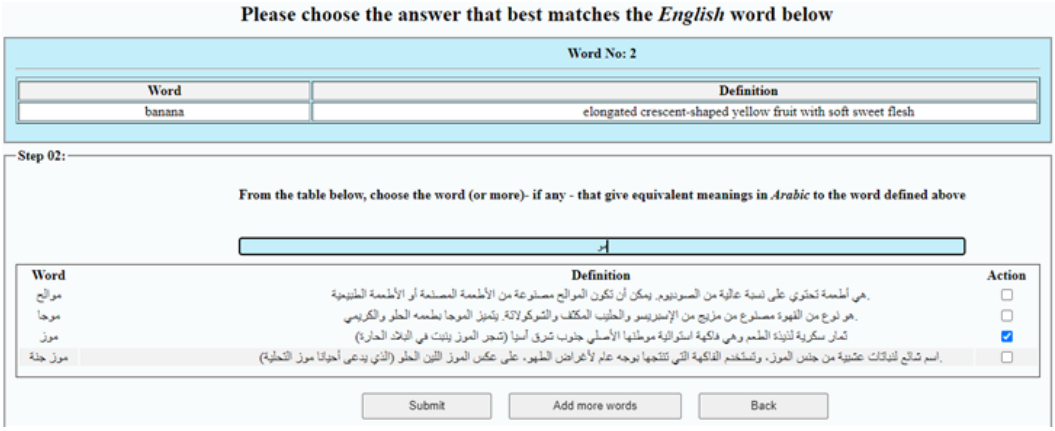


Figure 9.4: Worker's GUI showing *Step 2* in using LingoGap.

answers from w'_1 were disregarded due to a failed ACQ question. Subsequently, answers from w'_2 and w'_3, with the Alpha value of 0.922, were used (see Table 10.6). Additionally, outlier responses were identified based on the average completion time for answering each English word. For instance, among 68 tasks, 19 answers were omitted because their completion time ranged from 3 to 6 seconds, whereas the average completion time for all tasks ranged from 80 to 120 seconds.

Task validation

We employed the methodology depicted in Figure 7.5 to validate crowdsourced data from G_1, G_2, G_3, and G_4. The agreement for tasks assigned to the groups surpassed the threshold (0.70), except for Task 5 by G_2 and Task 17 by G_3. Consequently, Task 5 was reallocated to workers from G_1 and G_2, resulting in an increase in Alpha from 0.59 to 0.89. Where, w'_3 from G_2 was replaced in subsequent tasks (6 and 7) due to inaccuracies in his results for Task 5, which caused disagreements among workers. Similarly, Task 17 was reallocated to workers from G_3 and G_4, resulting in an increase in Alpha from 0.62 to 0.82. Where, w'_1 from G_3 was replaced in subsequent tasks (18 to 27) due to inaccuracies in his results for Task 17, which again led to disagreements among workers.

Secondly, all new Arabic words (missing from the Arabic input dataset) and non-100% IAA responses (for 88 English words) for each task underwent expert validation. The expert confirmed all new words and proposed alternatives for two gaps. For instance, خبز شراك "*khubz shrak*" was suggested for "*chapatti*" instead of leaving a gap, as initially suggested by G_3 in Task 12. In total, 1532 gaps, 832 equivalent words, and 100 new Arabic words were identified, with an average Alpha score of 0.84. More details on the crowdsourced Arabic data are shown in Table 10.6 and Table 10.7.

9.1.2 Arabic → English

Task generation

This experiment was also conducted within the food domain, but in the reverse direction compared to Experiment 1, as illustrated in Figure 7.2, exploring the diversity aspects from Arabic to English. We used the same materials prepared during the task generation phase of the previous experiment, including the English food and Arabic food datasets previously created. However, for the Arabic dataset, we eliminated the Arabic words (and their definitions) that overlapped and were identified as equivalent to English words in Experiment 1 (English-to-Arabic). In total, after eliminating overlaps (and the Arabic words duplicated in the overlap set), 906 Arabic words (from an original list of 1,607) were used in this experiment.

Crowd selection

Thirteen native English speakers with a good command of Arabic were recruited from the United Kingdom and the United States. They are university-educated, with different majors such as medicine, nutrition, business administration, linguistics, and computer engineering. Twelve of them met the criteria for high-quality crowd selection and participated as volunteers. Moreover, we applied the same distribution of members as in Experiment 1—splitting them into four groups—G_1, G_2, G_3, and G_4—each consisting of three members.

Contribution collection and quality control

We used the Arabic food dataset, containing 906 words, in our experiment. Each participant group, G_1 and G_2, compared approximately 245 Arabic food words (G_1 tasked with 245 words and G_2 with 241 words) with the complete English food dataset across seven crowdsourcing tasks. Similarly, each group of G_3 and G_4, compared 210 Arabic food words with the complete English food dataset across six tasks. G_1 and G_2 focused on *sweets, rice meals, soups, vegetables, meats, sandwiches*, while G_3 and G_4 addressed *drinks, pizza, salads, dairy products, bread,* and *fruits*. For the concerns of *pilot tasks*, we used the same parameters shown in Table 9.1. Here, the selected crowd conducted four pilot tasks. As a result of the pilot tasks in Experiment 1, we opted for 35 questions, 3 ACQs, and set up 60 minutes for each task.

We used LingoGap's admin GUI to generate crowdsourcing tasks, which involved providing task descriptions, selecting Arabic as the source language, English as the target language, and choosing 35 Arabic words from the dataset. Furthermore, we incorporated experiment guidelines.

Crowd workers reviewed the task guidelines (which are described in Table 9.2) upon logging into LingoGap. For each Arabic word, LingoGap presents three steps: *Step 1* asks if there's an equivalent English word; if confirmed, it proceeds, otherwise, it records a lexical gap. *Step 2* prompts selecting an equivalent English word(s); if not found, *Step 3* permits adding equivalents in textboxes.

Just like in Experiment 1, three ACQs were employed for real-time quality control of crowdsourced gaps and words in each task. Furthermore, the average completion time for answering each Arabic word aided in identifying outlier responses. Table 10.8 illustrates crowdsourced data per task, encompassing gap count, new English words (missing from the English input dataset), and worker's agreement (Alpha).

Task validation

We employed the methodology depicted in Figure 7.5 to validate crowdsourced data from the groups G_1, G_2, G_3 and G_4. Alpha values for all tasks of the groups surpassed the threshold (0.70), as depicted in Table 10.8. Nevertheless, all new English words and answers (for 79 Arabic words) with less than 100% IAA were forwarded for expert validation. The expert confirmed all new English words. He also identified four Arabic words as gaps, for instance, مهلبية, meaning "*a type of dessert made from milk, starch, and sugar*", identified as a lexical gap, rather than the initially provided (by G_2 in in Task 4) equivalent, *pudding*, which means "*any of various soft sweet desserts thickened usually with flour and baked or boiled*". Consequently, we identified 608 gaps in English, 298 equivalent words, and 49 new English words with an average Alpha score of 0.85. More details on the crowdsourced English data are shown in Table 10.8.

9.2 Experiment 2 (Indonesian vs Banjarese: food)

9.2.1 Indonesian → Banjarese

Task generation

For this experiment, we conducted a crowdsourcing task in the food domain from Indonesian to Banjarese. To achieve this purpose, we created two datasets: Indonesian dataset that consists of 1,448 terms, and Banjarese dataset that consists of 812 terms. Both datasets were obtained by using semantic filtering methods. For the Indonesian dataset, we used similar methods (Figures 9.2 and 9.3) as we did for Arabic. This method consists of three steps: in *Step 1*, we used the Kamus Bahasa Indonesia [146] to find all terms in Indonesian. In *Step 2*, we used IndoBERT [88] to transform all terms from Step 1 into vectors. In *Step 3*, the vectors are clustered to the term "food", and we used cosine similarity to find terms that are closest to

the centroid. A custom Python script based on IndoBERT was implemented following the methodology to collect food-related words in Indonesian from the the dictionary of Kamus Bahasa Indonesia, see the script uploaded to GitHub[2].

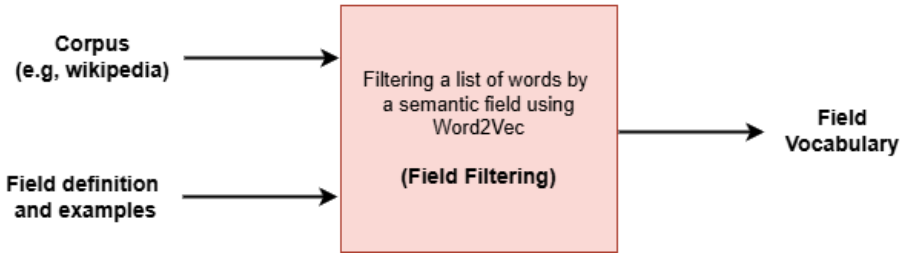


Figure 9.5: Semantic field retrieving method for Banjarese.

In the case of the Banjarese dataset, the absence of a high-quality dictionary or a monolingual language model posed a significant challenge. To address this, we utilized alternative resources, including Word2Vec [105], Wiktionary, and Wikipedia, to compile the dataset through a three-step process. Figure 9.5 provides a general overview of the methodology, while Figure 9.6 presents its flowchart with examples, detailing the inputs, processing steps, and outputs. For clarity and ease of understanding, these examples were provided in *English*. In *Step 1*, we sourced a Banjarese corpus by extracting data from the NLLB [144] corpus via NusaCrowd [33]. This corpus was then used to build and train a static word embedding model using Word2Vec. In *Step 2*, we employed a custom Python script[3] to extract a list of terms and their definitions from both Wiktionary and Wikipedia. These terms were transformed into vector representations using the Word2Vec model developed in Step 1. In *Step 3*, the vectorized terms were clustered, with the term "food" and its definition serving as the centroid. Cosine similarity was utilized to calculate the distance between each term and the centroid, with a threshold set at 0.85. Any terms exceeding this threshold were classified as food-related and subsequently added to the dataset.

Crowd selection

In this experiment, we recruited six undergraduate students who are native Banjarese speakers with good proficiency in Indonesian. These students are currently pursuing their bachelor's degrees at a university where Banjarese is the native language, ensuring their linguistic expertise. All participants met

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

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

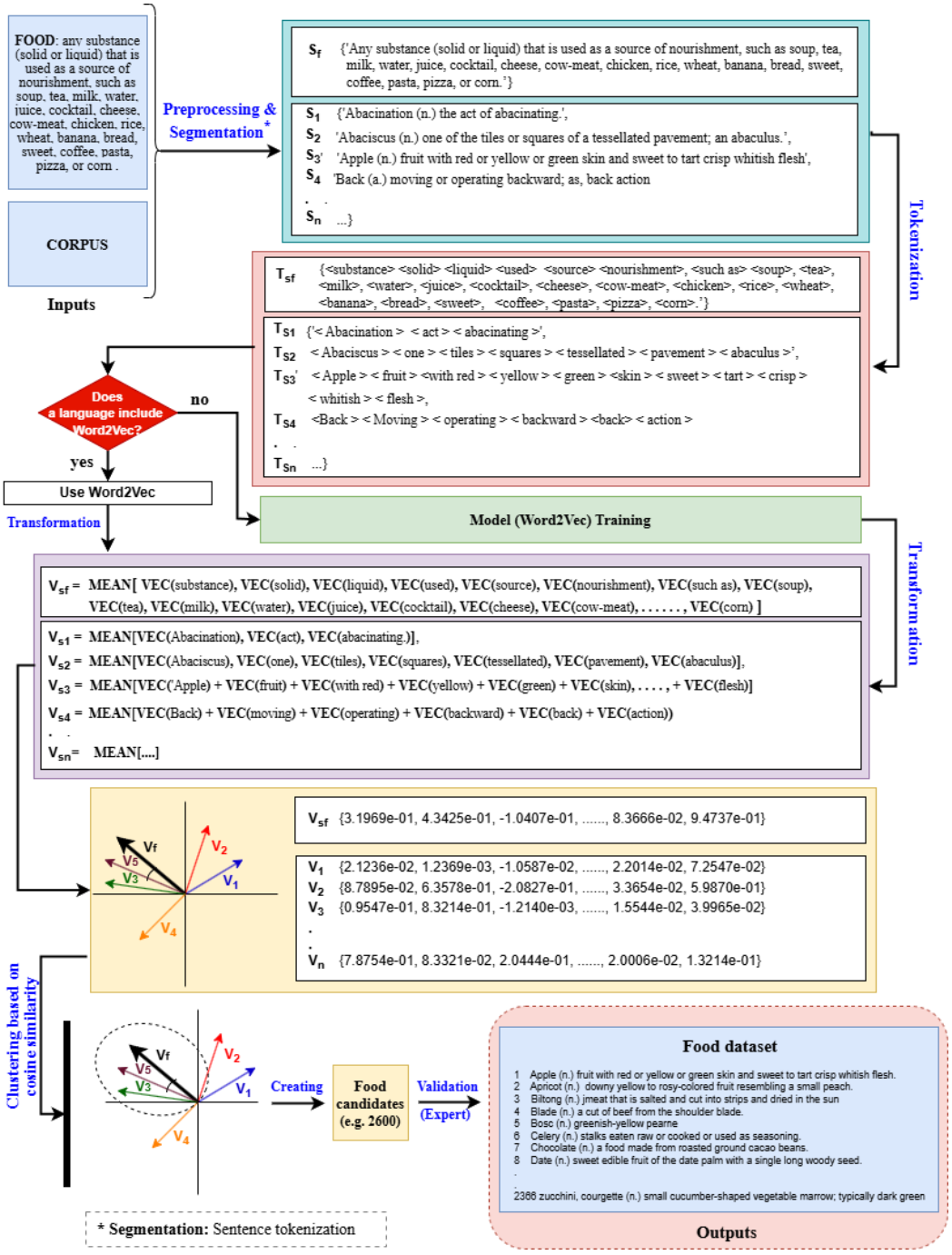


Figure 9.6: Flowchart of the semantic-field filtering method to collect food words from a corpus (e.g., Wikipedia).

the criteria for high-quality crowd selection. To facilitate the experiment,

the students were divided into two groups, each consisting of three members.

Contribution collection and quality control

The volunteers proceeded to carry out the assigned tasks following the predefined instructions and guidelines. The process mirrored previous pilot experiments, with one key distinction: this time, Indonesian and Banjarese terms (instead of English and Arabic) were used. The terms were randomly distributed, ensuring that no group focused on specific food categories. During each session, the volunteers were tasked with matching 45 Indonesian terms to their Banjarese equivalents. To facilitate this, we utilized LingoGap's admin GUI, where we defined Indonesian as the source language, Banjarese as the target language, and provided the set of 45 Indonesian terms per task for comparison.

In each session, volunteers began by logging into LingoGap using their assigned identifier. They were then prompted to review the provided guidelines, with the option to accept or reject them. Upon acceptance, LingoGap presented an Indonesian word along with its definition, followed by the question: "Does the Banjarese language have an equivalent for the word above?" Volunteers were given three response options: "Yes", "No", or "I don't know". If they selected "No" or "I don't know", a new Indonesian word (to be investigated) and definition were displayed. If they answered "Yes", a checklist of Banjarese words and definitions appeared, offering the options to "Submit", "Add more words", or "Back".

Volunteers were required to select at least one word from the list to submit. However, if they believed none of the provided words matched the Indonesian term, they could choose "Add more words", allowing them to manually input the correct Banjarese word and its definition. This process continued until the Indonesian word list for the session was exhausted, at which point the session concluded. The results of this experiment are presented in Table 10.9.

Task validation

To ensure the reliability of the volunteers' responses, we implemented the validation methodology illustrated in Figure 7.5. A minimum agreement threshold of 0.7 was established, and both groups consistently exceeded this value. Overall, the experiment yielded 750 lexical gaps, 507 equivalent words, and 43 new words (absent from the original Banjarese input dataset) in Banjarese, with an average Alpha score of 0.83, demonstrating strong consensus among participants.

9.2.2 Banjarese → Indonesian

Task generation

This experiment was also conducted within the food domain, but in the reverse direction of the previous experiment, exploring linguistic diversity from Banjarese to Indonesian. The same materials prepared during the task generation phase of the previous experiment were utilized, with some modifications. Specifically, the terms in Banjarese that had equivalent terms in Indonesian, as identified in the previous experiment, were excluded. As a result, the current experiment uses 330 Banjarese terms out of the original 812.

Crowd selection

Consistent with our previous experiment, we recruited six volunteers, all of whom are native Indonesian speakers with good proficiency in Banjarese. Each volunteer successfully passed the quality control criteria for high-quality crowd selection. The volunteers were then divided into two groups, each consisting of three members, to facilitate the task.

Contribution collection and quality control

This experiment employed a methodology consistent with the previous one, utilizing the same dataset of food-related terms, with Banjarese as the source language and Indonesian as the target language. As in the prior experiment, the terms were randomly distributed among the volunteers. For this iteration, we used LingoGap's admin interface to create the task, specifying Banjarese as the source and Indonesian as the target language. Volunteers were assigned 45 terms (as observed in the pilot experiment) per task to work on. The results of this experiment are presented in Table 10.10.

Task validation

Similar to the previous experiment, here we also utilized the methodology depicted in Figure 7.5 to ensure quality, with minimum agreement threshold at 0.7. In this case, both groups also exceed the threshold. However, there are some entries without 100% inter-annotator agreement that were submitted to the expert. As an example, for the word *rabuk*, meaning "*a type of grinded meat*" in Banjarese, some participants filled it in as a gap. This is corrected by the expert with the word *abon* in Indonesian. This experiment results in 201 gaps, 98 equivalent words, and 30 new Indonesian words (missing from the original Indonesian input dataset) in Indonesian, with average Alpha score of 0.83.

9.3 Experiment 3 (English vs. Persian: kinship)

Persian (Farsi) is the official language of Iran and is spoken in 28 other countries, with 78 million native speakers and approximately 130 million total speakers, including second-language speakers (e.g., in Uzbekistan). It belongs to the Indo-European language family and originates from Eurasia. Iran, where Persian is predominantly spoken, is linguistically diverse, with 76 languages in use. Despite this diversity, Persian functions as the standard language.

This study investigates linguistic diversity in kinship terminology between Persian and English through two distinct experiments: the first examines the diversity from English to Persian, while the second focuses on the reverse direction, from Persian to English. The crowdsourcing methodology described in Chapter 7 was employed in both experiments. The findings are described below.

9.3.1 English → Persian

Task generation

We conducted this experiment on the kinship domain, crowdsourcing from English to Persian. Two input kinship datasets were created: the first was the English dataset, which included 17 terms, and the second was the Persian dataset, containing 38 terms. The English kinship terms were collected from the UKC using the Exporter tool, following the same methodology used in Section 9.1 for constructing the English food terms, but, in this experiment, we use the "kinship" term as the representing concept for the kinship semantic field inserted to the tool GUI, as illustrated in Figure 9.1.

FarsNet [108], the Persian WordNet, is one of the most well-known and high-quality linguistic resources for Persian, containing over 100,000 words. Consequently, we utilized it to construct the Persian kinship input dataset. To achieve this, we developed a custom Python script that follows the methodology outlined in Figure 9.7 to extract Persian kinship terms from FarsNet. The script takes one or more synsets (e.g., خویشاوندی, meaning "*kinship*") representing the kinship domain, along with the FarsNet resource, as inputs. It then retrieves all lexical units associated with the specified synsets from FarsNet. See the script uploaded to GitHub[4].

Crowd selection

In this experiment, we recruited three undergraduate students (as volunteers) who are native Persian speakers with strong proficiency in English. These students are currently pursuing bachelor's degrees at universities

[4] https://github.com/HadiPTUK/developed_scripts/blob/main/persian_kinship_terms_script.py

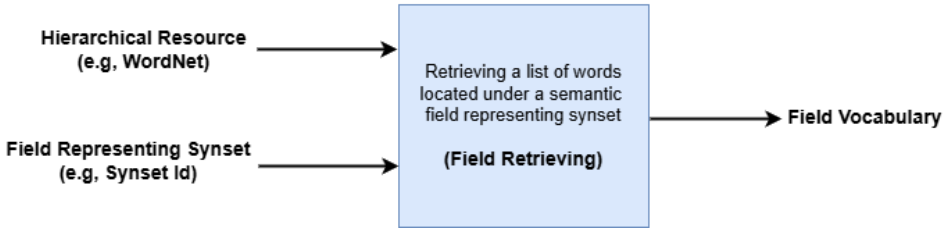


Figure 9.7: Semantic field retrieving method for Persian.

where Persian is the native language, ensuring their linguistic expertise. All participants met the criteria for high-quality crowd selection.

Contribution collection and quality control

This experiment utilized the crowdsourcing methodology outlined in Chapter 7, employing the English and Persian kinship datasets, with English as the source language and Persian as the target language. As in the previous experiments, the terms were randomly distributed among the volunteers. For this iteration, we used LingoGap's admin interface to create the task, specifying English as the source and Persian as the target language. Volunteers were assigned a single task comprising all 17 English kinship terms, including two ACQs, with a duration of 45 minutes.

Task validation

We applied the methodology outlined in Figure 7.5, using a minimum agreement threshold of 0.7 to ensure the quality of the collected data. In this phase, since there was only one group of crowd workers, we incorporated assistance from the group recruited for the subsequent experiment (from Persian to English). Both groups exceeded the agreement threshold. However, no entries were submitted to an expert, as the inter-annotator agreement for each contribution was 100%. As a result, this experiment yielded nine gaps and eight equivalent words, with an Alpha score of 1.

9.3.2 Persian → English

Task generation

This experiment was also conducted within the kinship domain, but in the reverse direction of the previous experiment, exploring linguistic diversity from Persian to English. The same materials prepared during the task generation phase of the previous experiment were utilized, with some modifications. Specifically, the terms in Persian that had equivalent terms (eight terms) in English, as identified in the previous experiment, were excluded. As a result, the current experiment uses 30 terms out of the original 38.

Crowd selection

In this experiment, we recruited three volunteers (students), all of whom are native English speakers with good proficiency in Persian. Each volunteer successfully passed the quality control criteria for high-quality crowd selection.

Contribution collection and quality control

This experiment employed a methodology consistent with the previous one, utilizing the same dataset of kinship-related terms, with Persian as the source language and English as the target language. As in the English-to-Persian experiment, the terms were randomly distributed among the volunteers through LingoGap. The LingoGap admin interface was used to create the task, specifying Persian as the source and English as the target language. Volunteers were assigned a single task comprising all 30 Persian kinship terms, including three ACQs, with a duration of 60 minutes.

Task validation

Similar to the previous experiment, this study employed the methodology depicted in Figure 7.5 to ensure the quality of collected contributions, with a minimum agreement threshold of 0.7. Since there was only one group of crowd workers in this case, assistance was provided by the group recruited for the previous experiment (English-to-Persian). Both groups exceeded the threshold. However, some entries without 100% inter-annotator agreement were submitted to an expert for review. For instance, the Persian term دایی meaning "maternal uncle", was misclassified by some participants, who provided the English term "uncle" as an equivalent meaning. The expert corrected this by identifying it as a lexical gap in English. This experiment ultimately identified 27 lexical gaps, three equivalent words, and three new words in English.

9.4 Experiment 4 (Arabic vs. Persian: kinship)

9.4.1 Arabic → Persian

Task generation

We conducted this experiment in the kinship domain, crowdsourcing diverse data from Arabic to Persian. Two input kinship datasets were utilized: the first was an Arabic dataset comprising 38 terms, and the second was a Persian dataset containing 38 terms. The Arabic kinship terms were generated using AraBERT and the Almaany dictionary, following the methodology

outlined in Figure 9.3 for constructing Arabic food terms. In this experiment, however, we used the terms صلة قرابة and قرابة, both meaning "*kinship*", along with examples of family relationships such as أب "*father*", أم "*mother*", أخ "*brother*", and أخت "*sister*". These terms served as the corpus for the kinship semantic field applied to the methodology. For the Persian kinship dataset, we used the Persian input dataset developed from the Farsnet lexicon during the previous English-Persian experiment.

Crowd selection

In this experiment, we recruited three undergraduate student volunteers who are native Persian speakers with strong proficiency in Arabic. These students are currently pursuing bachelor's degrees at universities where Persian is the primary language, ensuring their linguistic expertise. All participants met the criteria for high-quality crowd selection.

Contribution collection and quality control

This experiment employed the crowdsourcing methodology described in Chapter 7, utilizing Arabic and Persian kinship datasets, with Arabic as the source language and Persian as the target language. Consistent with previous experiments, the terms were randomly distributed among the volunteers using the LingoGap platform. The LingoGap admin interface was used to create a task, specifying Arabic as the source and Persian as the target language. Volunteers were assigned two tasks, each containing 19 Arabic kinship terms and 2 AcQs, with a duration of 45 minutes. The results of this experiment are presented in Table 9.3.

Table 9.3: Information on the crowdsourced data in Persian.

	Gaps	**Words**	**New Concepts**	**Alpha**
Task 1	**6**	**13**	**0**	**0.85**
Task 2	**10**	**9**	**0**	**0.73**
Total	**16**	**22**	**0**	**0.79 (avg)**

Task validation

We applied the methodology outlined in Figure 7.5, setting a minimum agreement threshold of 0.7 to ensure the quality of the collected data. During this phase, since only one group of crowdworkers was available, we incorporated assistance from the group recruited for the subsequent experiment (Persian-to-Arabic). Both groups exceeded the agreement threshold. However, some entries that did not achieve 100% inter-annotator agreement

were submitted to an expert for review. For instance, the Arabic term حفيد, meaning "*grandson*", was mismatched by two participants, who provided the Persian term نوه, meaning "*grandchild*", as an equivalent. The expert resolved this by identifying it as a lexical gap in Persian. As a result, this experiment identified 16 lexical gaps and 22 equivalent words in Persian, achieving an Alpha score of 0.79. Additionally, no new words were identified in this experiment.

9.4.2 Persian → Arabic

Task generation

This experiment was also conducted within the kinship domain, but in the reverse direction of the previous experiment, exploring linguistic diversity from Persian to Arabic. The same materials prepared during the task generation phase of the previous experiment were utilized, with some modifications. Specifically, we excluded the terms in Persian that had equivalent terms (22 terms) in Arabic, as identified in the previous experiment. As a result, the current experiment uses 16 terms out of the original 38 Persian terms.

Crowd selection

In this experiment, we recruited three volunteers (students), all of whom are native Arabic speakers with good proficiency in Persian. Each volunteer successfully passed the quality control criteria for high-quality crowd selection.

Contribution collection and quality control

This experiment employed a methodology consistent with the previous one, utilizing the same dataset of kinship-related terms, with Persian as the source language and Arabic as the target language. As in the previous experiment, the terms were randomly distributed among the volunteers through LingoGap. The LingoGap admin interface was used to create the task, specifying Persian as the source and Arabic as the target language. Volunteers were assigned a single task comprising all 16 Persian kinship terms, including two ACQs, with a duration of 45 minutes.

Task validation

Similar to the previous experiment, this validation followed the methodology illustrated in Figure 7.5, applying a minimum agreement threshold of 0.7 to ensure the quality of the collected contributions. As only one group of crowd workers participated in this case, assistance was provided by the group

recruited for the previous experiment (Arabic-to-Persian). Both groups exceeded the threshold. However, some entries that did not achieve 100% inter-annotator agreement were reviewed by an expert. As a result, this experiment identified 13 lexical gaps and three equivalent words in Arabic.

9.5 Experiment 5 (English vs. Turkish: kinship)

Turkish, a member of the Oghuz branch of the Turkic language family, is spoken by approximately 90 million native speakers. It is the official language of Turkey and one of the recognized languages of Cyprus. Turkish also exerts linguistic influence across the Balkans, Central Asia, the Middle East, and among Turkish-speaking communities throughout Europe.

This study investigates linguistic diversity in kinship terminology between Turkish and English through two distinct experiments: the first examines the diversity from English to Turkish, while the second focuses on the reverse direction, from Turkish to English. The crowdsourcing methodology described in Chapter 7 was employed in both experiments. The findings are described below.

9.5.1 English → Turkish

Task generation

We conducted this experiment on the kinship domain, crowdsourcing from English to Turkish. Two input kinship datasets were created: the first was the English dataset, which included 17 terms, and the second was the Turkish dataset, containing 38 terms. We used the same English kinship dataset created in the English-Persian experiment in Section 9.3, which was imported from the UKC using the Exporter tool.

KeNet [14], the Turkish WordNet, a high-quality linguistic resource for Turkish comprises 76,757 synsets and 153,514 words. Consequently, we utilized it to construct the Turkish kinship input dataset. To achieve this, we developed a custom Python script that follows the methodology outlined in Figure 9.7 to extract Turkish kinship terms from KeNet. The script takes one or more synsets (e.g., *Akraba*, meaning "*kinship*") representing the kinship domain, along with the KeNet resource, as inputs. It then retrieves all lexical units associated with the specified synsets from KeNet. See the script uploaded to GitHub[5].

[5] https://github.com/HadiPTUK/developed_scripts/blob/main/turkish_kinship_terms_script.py

Crowd selection

Four native Turkish speakers with high proficiency in English were recruited as crowd workers. The participants were undergraduate students from diverse academic disciplines, including mechanical engineering, industrial engineering, computer engineering, and marketing. During the high-quality crowd selection process, one participant was replaced following Alpha filtering.

Contribution collection and quality control

This experiment employed the crowdsourcing methodology described in Chapter 7, utilizing English and Turkish kinship datasets, with English as the source language and Turkish as the target language. Consistent with previous experiments, the source terms were randomly distributed among the volunteers using the LingoGap platform. The LingoGap admin interface was used to create the task, specifying English as the source and Turkish as the target language. Volunteers were assigned a single task comprising all 17 English kinship terms, including two ACQs, with a duration of 45 minutes.

Task validation

We applied the Alpha filtering algorithm outlined in Figure 7.5, setting a minimum agreement threshold of 0.7 to ensure the high quality of the collected data. During the validation, as only one group of crowdworkers was available, we incorporated assistance from the group recruited for the subsequent experiment (Turkish-to-English). Both groups exceeded the agreement threshold. However, one English word, "sister", which did not achieve 100% inter-annotator agreement, was submitted to an expert for review. For instance, while three participants identified it as a lexical gap, the expert corrected it by providing the Turkish equivalent, "kız kardeş", As a result, this experiment identified six lexical gaps and 11 equivalent words in Turkish, achieving an Alpha score of 0.94. Furthermore, no new words were identified in this experiment.

9.5.2 Turkish → English

Task generation

This experiment was also conducted within the kinship domain, but in the reverse direction of the previous experiment, exploring linguistic diversity from Turkish to English. The same materials prepared during the task generation phase of the previous experiment were utilized, with some modifications. Specifically, the terms in Turkish that had equivalent terms (11 terms)

in English, as identified in the previous experiment, were excluded. As a result, the current experiment uses 27 terms out of the original 38.

Crowd selection

In this experiment, we recruited three volunteers (students), all of whom are native English speakers with good proficiency in Turkish. Each volunteer successfully passed the quality control criteria for high-quality crowd selection.

Contribution collection and quality control

This experiment employed a methodology consistent with the previous one, utilizing the same dataset of kinship-related terms, with Turkish as the source language and English as the target language. As in the English-to-Turkish experiment, the terms were randomly distributed among the volunteers through LingoGap. The LingoGap admin interface was used to create the task, specifying Turkish as the source and English as the target language. Volunteers were assigned a single task comprising all 27 Turkish kinship terms, including three ACQs, with a duration of 60 minutes.

Task validation

Similar to the previous experiment, this study employed the Alpha filtering methodology illustrated in Figure 7.5 to ensure the high quality of the collected contributions, with a minimum agreement threshold of 0.7. As only one group of crowd workers was available in this case, assistance was provided by the group recruited for the previous experiment (English-to-Turkish). Both groups exceeded the threshold. However, some entries that did not achieve 100% inter-annotator agreement were submitted to an expert for review. For example, the Turkish term *oğul* meaning "*son*", was misclassified by three participants, who provided "child" as the equivalent meaning in English. The expert corrected this by confirming "son" as the appropriate English equivalent. This experiment ultimately identified 24 lexical gaps, three equivalent words, and one new word in English.

9.6 Experiment 6 (Kinship concepts vs. Turkish)

This experiment is a monolingual study that explores linguistic diversity by comparing language-independent kinship concepts with Turkish kinship terms. As described in Section 3.4, such concepts can be retrieved from the supra-lingual (top) layer of the UKC (see Figure 3.1). In this study, during the first experiment, we use the kinship concepts as source terms and Turkish kinship words as target terms. After completing the first experiment, we

identify overlaps—equivalent Turkish terms—and then determine Turkish terms (non-overlapping terms from the Turkish dataset) that do not have equivalents. Subsequently, we conduct a new experiment using these Turkish terms as source terms and the kinship concepts as target terms. The second experiment focuses on identifying new kinship concepts found in Turkish, specifically those that are unique to Turkish and not included in the original dataset of kinship concepts.

Kinship concepts → Turkish

Task generation

The same materials developed for the experiment investigating linguistic diversity across Arabic dialects and Indonesian languages, as described in Section 5.1 and Section 5.2, were used in this study. First, kinship concepts were formalized using the UKC's concept hierarchy, which encompasses the following subdomains: grandparents, grandchildren, siblings, uncles and aunts, nephews and nieces, and cousins. The categorization of brotherhood is depicted in the top layer of the UKC, as shown in Figure 3.1. Second, the dataset of kinship concepts was utilized, comprising 184 original kinship concepts and new ones (the only new concepts added to the UKC resource) identified from Arabic dialects and Indonesian languages, resulting in a total of 197 concepts. Table 5.1 presents the count of kinship concepts within each subdomain. As introduced in Section 5.1, this dataset consists of two columns: the concept itself and its definition, both provided in English.

For the Turkish kinship dataset, which contains 38 terms, we used the same dataset constructed in the English-Turkish experiment, as demonstrated in Section 9.5.

Crowd selection

Four native Turkish speakers, who conducted the English-to-Turkish experiment described in Section 9.5.2, participated in this experiment.

Contribution collection and quality control

In this experiment, we employed the crowdsourcing methodology described in Chapter 7, along with the kinship concepts and Turkish kinship datasets. The kinship concepts served as the source terms, while Turkish kinship words were the target terms. As in previous experiments, we distributed the source terms to volunteers using the LingoGap platform. The LingoGap admin interface was utilized to create the crowdsourcing tasks, specifying the kinship concepts as source terms and the Turkish words as target terms for each task. Volunteers were assigned seven crowdsourcing tasks, each focusing on the kinship concepts of a specific subdomain, except for the

"cousins" subdomain, for which two tasks were conducted due to its large number of concepts (56 in total).

For example, Task 1, conducted on the "grandparents" subdomain, included 18 source words, corresponding to the 18 kinship concepts in this subdomain. Similarly, Task 4, performed on the "uncles and aunts" subdomain, included 26 source words, matching the 26 kinship concepts it contains. Overall, this experiment identified 164 lexical gaps, 33 equivalent Turkish words, and four new kinship concepts in Turkish.

Task validation

The Alpha filtering methodology illustrated in Figure 7.5 was also used to ensure the high quality of the collected Turkish lexicalizations and gaps, with a minimum agreement threshold of 0.7. As only one group of crowd workers was available in this case, assistance was provided by the group recruited for the experiment of (Turkish to English). Both groups exceeded the threshold. However, some entries that did not achieve 100% inter-annotator agreement were submitted to an expert for review. For example, three participants provided the Turkish terms ***dede, büyükbaba*** meaning "*grandfather*", as equivalents to the concept "*paternal grandfather*", while the fourth participant identified it as a lexical gap. The expert corrected this by confirming the concept *paternal grandfather* as a gap in Turkish.

Turkish → Kinship concepts

This experiment focuses on the new kinship concepts identified in Turkish. We invited the same four native English speakers who participated in the Turkish-to-English experiment described in Section 9.5.2 to take part in this experiment. In a single crowdsourcing task conducted using LingoGap, these participants compared *four* concepts identified in the previous experiment with the full list of kinship concepts (197 concepts). The crowdsourced contributions were validated using the Alpha Filtering Methodology shown in Figure 7.5. The results indicated that the inter-participant agreement (Alpha) for the four concepts was 1.0, confirming these concepts.

9.7 Experiment 7 (Kinship concepts vs. Persian)

This experiment is a monolingual study aimed at exploring linguistic diversity by comparing language-independent kinship concepts with Persian kinship terms. We employed the crowdsourcing methodology (Chapter 7) to conduct this study. As in the previous experiment (Section 9.6), the UKC lexicon was utilized for the formalization of kinship concepts using the kinship hierarchy and for preparing the input dataset of kinship concepts. The details of the experiment are described below.

Kinship concepts → Persian

Task generation

We prepared two input datasets: the first dataset comprises 197 kinship concepts developed during a previous experiment comparing kinship concepts with Turkish (see Section 9.6), while the second dataset includes 38 Persian kinship terms constructed during an experiment comparing English kinship terms with Persian (see Section 9.3). Additionally, in this phase, we formalized the kinship concepts using the UKC's concept hierarchy, encompassing the same kinship subdomains adopted in the previous experiment.

Crowd selection

Three native Persian speakers, who conducted the English-to-Persian experiment described in Section 9.3.1, participated in this experiment.

Contribution collection and quality control

As in the previous experiment, we distributed the source terms to volunteers using the LingoGap platform. The LingoGap admin interface was used to create crowdsourcing tasks, specifying the kinship concepts as source terms and the Persian words as target terms for each task. Seven crowdsourcing tasks were assigned to participants, following the same procedure as in the previous experiment. Each task focused on the kinship concepts of a specific subdomain. For example, Task 2, which addressed the "grandchildren" subdomain, included 27 source words corresponding to the 27 kinship concepts within this subdomain.

Task validation

To ensure the high quality of the collected Persian words and lexical gaps, we applied the Alpha filtering methodology depicted in Figure 7.5. Since only one group of crowd workers conducted this experiment, we requested a second group of volunteers, originally recruited for the Persian-to-English experiment, to participate in the validation process. Both groups exceeded the predetermined minimum threshold of Alpha (0.7).

However, entries that did not achieve 100% inter-annotator agreement were submitted for expert review. For instance, two participants identified the concept "*breastfeeding brother*" as a lexical gap in Persian. Upon review, the expert corrected this by providing the Persian word برادر رضاعی as an equivalent for this concept.

9.8 Experiment 8 (Kinship concepts vs. Maba)

The Maba language, a member of the Nilo-Saharan language family, is primarily spoken in eastern Chad, particularly in the Ouaddaï region, by the Maba people. With approximately 300,000 speakers, Maba serves as a crucial linguistic and cultural marker for the ethnic group, which has historically been influenced by Arab and Sudanese cultures due to trade, migration, and Islamic expansion. The language incorporates numerous Arabic loanwords, reflecting prolonged linguistic interaction, especially in religious and commercial contexts. Predominantly unwritten, Maba relies on oral tradition to transmit folklore, history, and social norms. Its distinct phonetic and grammatical structure is exemplified by common words such as *aŋ* "*water*" and *dŋ* "*fire*". Despite its regional significance, Maba faces challenges from the increasing dominance of Arabic and French, Chad's official languages, which exert growing influence on younger generations and formal education.

This experiment is a monolingual study aimed at exploring linguistic diversity by comparing language-independent kinship concepts with Maba kinship terms. We employed the crowdsourcing methodology (Chapter 7) to conduct this study. As in the previous two experiments (Sections 9.6 and 9.7), the UKC lexicon was utilized for the formalization of kinship concepts using the kinship hierarchy and for preparing the input dataset of kinship concepts. The details of the experiment are described below.

Kinship concepts → Maba

Task generation

We prepared two input datasets: the *first dataset* consists of 197 kinship concepts, developed and formalized using the UKC's concept hierarchy employed in the previous experiment that compared kinship concepts with Persian (see Section 9.7). The *second dataset* contains 20 Maba kinship terms, constructed using a *fieldwork-based data collection method* (illustrated in Figure 9.8), which involved gathering terms directly from native Maba speakers.

This method was designed to construct SF datasets for *undigitized languages* that have not been represented or encoded in digital formats. Such languages often lack online dictionaries, lexical resources, text corpora, or other digital materials, making it challenging for computational linguists to access linguistic data. The method consists of two scenarios: *Scenario 1* outlines the process of collecting field words from *printed textual resources* (e.g., a paper dictionary), while *Scenario 2* details the collection of field words directly from native speakers through *fieldwork*.

The choice of a scenario depends on the availability of printed resources in the studied language. If a paper resource is available, the requester collects field words from it by following the steps outlined in *Scenario 1* (depicted

on the left-hand side of the flowchart). This scenario involves digitizing the paper resource and subsequently applying the semantic filtering method using Word2Vec, as illustrated in Figure 9.6. To digitize a paper resource (e.g., a dictionary), the following steps are performed:

1. Scan the Pages: Each page of the dictionary is scanned using a suitable device (e.g., a computer scanner or smartphone). The scanned images should be clear and legible.

2. Organize the Images: The scanned images are stored in a designated folder and arranged logically, either alphabetically or numerically, based on the dictionary's structure.

3. Convert Images to Text: Optical Character Recognition (OCR) software is used to convert the scanned images into text. Various OCR tools, such as Adobe Acrobat, ABBYY FineReader, Tesseract OCR, and Google Drive, offer this functionality.

4. Proofread and Correct Errors: The extracted text is reviewed and corrected for any errors introduced by the OCR process. This step is crucial, as OCR may misinterpret certain characters or words, especially in cases of unusual fonts or formatting.

5. Save the Digital Text: The final corrected text is saved in plain text (TXT) format and given an appropriate name, such as Corpus C.

In this scenario, the digitized corpus (e.g., Corpus C), along with a list containing the definition of a semantic field and a set of example terms, is input into the method illustrated in Figure 9.6. The data is then processed following the steps outlined in Section 9.2 and ultimately contributes to the creation of the final dataset.

When a printed resource is unavailable in the target language—as is the case for Maba, which lacks paper resources, making SF word data collection particularly challenging—*Scenario 2* (depicted on the right-hand side of the flowchart) is employed. This scenario outlines the method for collecting field words from native Maba speakers through *fieldwork*.

To create the dataset of Maba kinship terms, the first step was identifying the kinship subdomains, which include grandparents, grandchildren, siblings, uncles and aunts, nephews and nieces, and cousins. A linguistic expert, who is both a native Maba speaker and a PhD student specializing in knowledge graphs, assisted in selecting a *seed word* for each subdomain (e.g., grandson for the grandchildren category). Subsequently, the expert recruited two additional *native Maba speakers* and instructed them to list words and their definitions related to each subdomain, using the seed word as a reference. Next, the collected data was organized into a structured dataset (e.g., a spreadsheet) with columns including the *word* provided by

the speaker, *collection date*, and *word occurrence*. Finally, the expert validated the dataset by correcting errors, removing redundant words and definitions, and storing the validated words as the list of kinship candidates in Maba.

Crowd selection

In this experiment, we recruited three student volunteers who are native Maba speakers with good proficiency in English. The group consisted of two master's students, who contributed to the construction of the Maba kinship dataset, and one undergraduate student. All participants completed their schooling and are currently pursuing their degrees at universities in Chad within a Maba-speaking community, ensuring their linguistic expertise. Furthermore, all participants met the criteria for high-quality crowd selection.

Contribution collection and quality control

As in the previous experiment, we distributed the source terms to volunteers using the LingoGap platform. The LingoGap admin interface was used to create crowdsourcing tasks, specifying the kinship concepts as source terms and the Maba kinship words as target terms for each task. Seven crowdsourcing tasks were assigned to participants, following the same procedure as described in Sections 9.6 and 9.7. Each task focused on the kinship concepts of a specific subdomain. For example, Task 1, which covered the "grandparents" subdomain, included 19 source words corresponding to the 19 kinship concepts within this subdomain.

Task validation

The Alpha filtering methodology, illustrated in Figure 7.5, was used to ensure the high quality of the collected Maba lexical terms and gaps, with a minimum agreement threshold of 0.7. Since only one group of crowd workers was available, a linguistic expert—a PhD student who assisted in constructing the Maba kinship terms during the task generation step—was asked to perform the validation, effectively acting as a second group of crowd workers.

Both groups exceeded the threshold; however, some entries that did not achieve 100% inter-annotator agreement were submitted for expert review. For instance, two participants provided the Maba term *mú ŋ́gák*, meaning "*paternal uncle*", as the equivalent of the concept "*uncle*", while the third participant identified it as a lexical gap. The expert resolved this discrepancy by confirming that the concept "*uncle*" is a lexical gap in Maba.

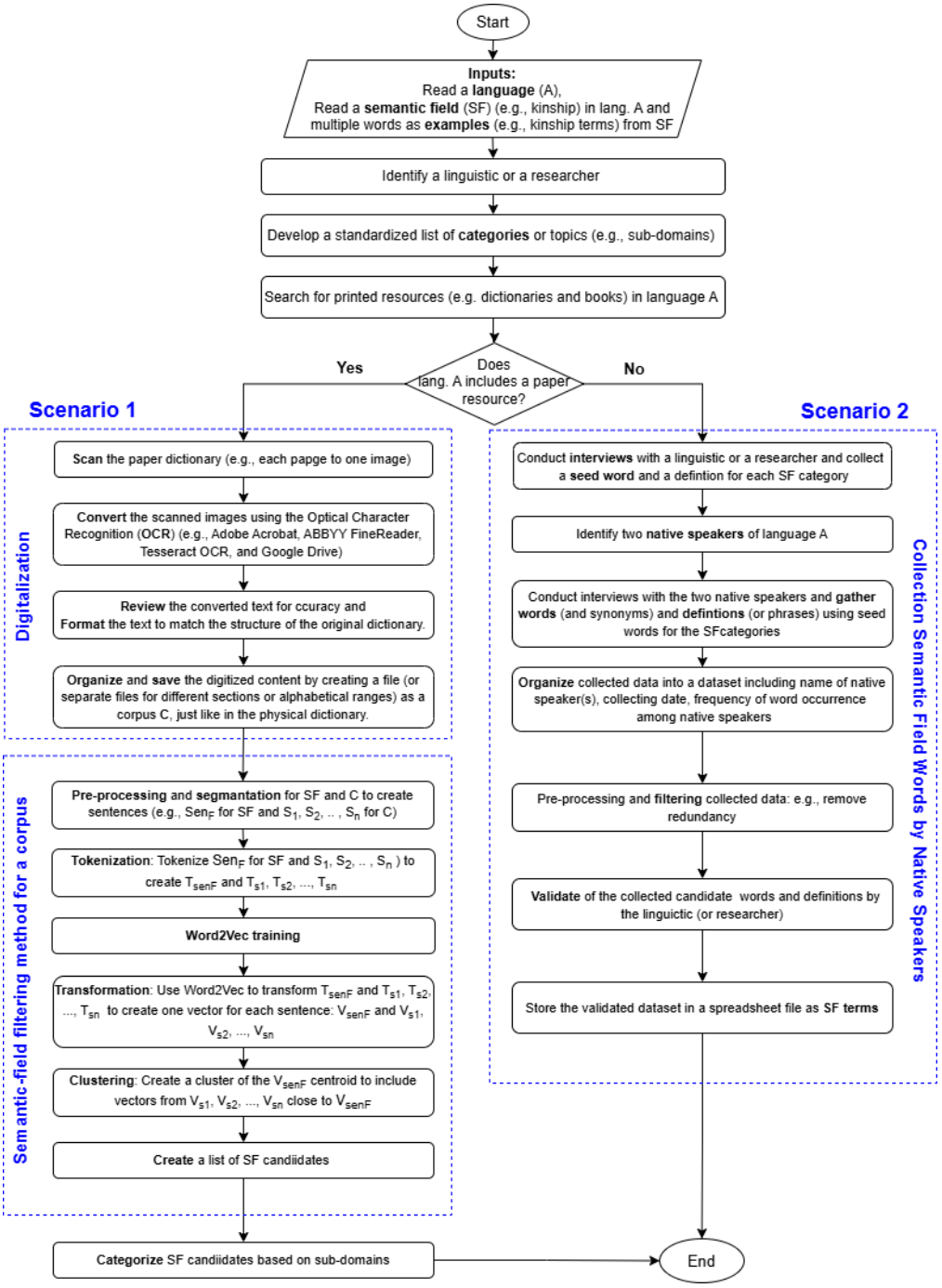


Figure 9.8: Flowchart of the data collection method for gathering semantic-field words (e.g., kinship terms) from undigitized languages.

9.9 LLMs as annotators

In this section, we explore the application of LLMs as annotation tools within linguistic research, focusing on the construction of small, diversity-aware datasets designed to benchmark the accuracy of diversity data obtained through our crowdsourcing methodology. To evaluate lexical diversity and meaning equivalence between source and target languages, we assess the translatability of lexical items, utilizing LLMs to capture nuanced cross-linguistic alignment.

Recent studies have shown that LLMs, particularly GPT-based models, are increasingly effective for text annotation tasks. Researchers such as Ding et al. [46] and Hasanain et al. [71] have employed GPT-3 and GPT-4 to annotate political tweets and detect propaganda in news articles, respectively, demonstrating both high quality and cost efficiency. Similarly, Sayeed et al. [135] utilizes Google's Gemini Pro language model to develop an annotation framework for extracting structured information from materials science literature. Comparative studies by Alizadeh et al. [5] and Törnberg [147] further suggest that GPT models often outperform crowdsourced annotators and even domain experts in tasks such as political affiliation identification. Building on these findings, this study explores the use of three LLMs—GPT-4, DeepSeek, and Gemini—for cross-linguistic annotation, aiming to analyze lexical diversity and meaning equivalence between languages. GPT-4 is specifically leveraged due to its superior performance over its predecessors.

In conducting the annotation process, we examine linguistic diversity across source and target language pairs (SL and TL) within a designated semantic field (SF). This process adheres to the crowdsourcing methodology detailed in Chapter 7, with two modifications. First, instead of the crowdsourcing step originally prescribed, we introduced a new phase termed the *Annotation Phase*, in which the LLM replaces crowdsourcing to perform annotation tasks. Each micro-task comprises 50 lexical entries (presented as a list) drawn from the SL within the specified SF, systematically paired with corresponding TL entries (inserted into a spreadsheet) as potential translation candidates. The primary objective here is to evaluate whether each SL entry possesses a translatable equivalent in the TL. Where an equivalent is identified, the corresponding TL entry is either selected from the dataset or proposed as a new term; where no equivalent exists, the SL entry is classified as a lexical gap in the TL. Figure 9.9 illustrates the template used to prompt the LLM for this assessment. Second, for validation, we relied exclusively on expert verification to evaluate and measure the accuracy of identified equivalent meanings and lexical gaps.

We applied this annotation method in two case studies within the food domain, investigating the *English-Arabic* and *Indonesian-Banjarese* language pairs. These language pairs were also analyzed using our crowdsourcing

"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 9.9: The template used to prompt each of the three LLMs—GPT-4, DeepSeek, and Gemini.

methodology, as described in Section 9.1 and Section 9.2, respectively. The first case study focused on *English and Arabic*, while the second examined *Indonesian and Banjarese*, as detailed below.

9.9.1 Case study 1: (English vs. Arabic: food terms)

In this case study, we conducted two experiments to explore diversity in both directions between English and Arabic, following the crowdsourcing approach. The first experiment focused on the lexicalization of English food terms into Arabic, resulting in a list of equivalent words and lexical gaps in Arabic. The second experiment reversed the direction, examining the lexicalization of Arabic food terms into English and producing a list of equivalents and gaps in English.

For the *English-to-Arabic* experiment, we utilized the datasets created in Section 9.1, selecting fifty English-specific food terms from a comprehensive English dataset of 2,364 words. These terms, along with the entire Arabic dataset of 1,607 words, were processed using GPT-4, DeepSeek, and Gemini. Each LLM employed zero-shot learning to annotate the English terms, either by identifying an equivalent Arabic word or marking the term as a lexical gap if no direct translation existed.

In the *Arabic-to-English* experiment, fifty Arabic-specific food terms were selected from the Arabic dataset. Both this subset and the full English dataset were processed using GPT-4, DeepSeek, and Gemini. Each model annotated the Arabic terms by either matching them with English equivalents or identifying lexical gaps.

To ensure accuracy, the identified Arabic words and gaps were validated

by a Ph.D. student specializing in lexical semantics and a native Arabic speaker. Meanwhile, the English words and gaps were reviewed by an Arabic university lecturer based in the United Kingdom. Table 9.4 presents the accuracy results of each model's annotations in both experiments, as assessed by the validators. *Accuracy* was defined as the number of correctly validated TL words or gaps, divided by the total number of SL terms. The validators observed two primary types of errors made by the LLMs in identifying lexical gaps: the model either provided an *incorrect TL* word or offered a *literal translation* of the SL term instead of correctly identifying a gap.

Exp. No	Source Language	Target Language	ChatGPT	DeepSeek	Gemini
1	English	Arabic	18	38	14
2	Arabic	English	42	46	38
3	Indonesian	Banjarese	8	10	36
4	Banjarese	Indonesian	40	46	58

Table 9.4: Percentage accuracy of data collected by ChatGPT-4, DeepSeek, and Gemini, as validated by experts.

For instance, in the *Arabic-to-English* experiment, the Arabic term الأبيضان "*water and yogurt*" was incorrectly translated as "*the two whites*" by GPT-4. Similarly, كبيبة "*a dish made of ground meat mixed with crushed rice or wheat, shaped into balls or patties, and cooked*" was literally translated as *kibbeh* by all three models, despite the term not existing in standard English lexicons.

In the *English-to-Arabic* experiment, the English term *malt liquor* meaning "*a lager with high alcohol content, legally distinct from beer*" was translated as مشروب شعير by GPT-4 and DeepSeek, meaning "*barley drink*" which inaccurately reflects the alcohol content. Similarly, the term *stout* meaning "*a strong, dark ale*" was rendered literally as ستاوت by Gemini and and DeepSeek, which does not exist in Arabic language resources.

9.9.2 Case study 2: (Indonesian vs. Banjarese: food terms)

Similar to the previous case study, this research focuses on the food domain, conducting two experiments between Indonesian and Banjarese.

In the first experiment, *Indonesian-to-Banjarese*, we utilized the datasets developed in Section 9.2. We selected fifty Indonesian-specific terms from a dataset of 1,448 Indonesian food-related words. These terms, along with the complete Banjarese dataset of 812 words, were processed by GPT-4, DeepSeek, and Gemini using zero-shot learning prompts. For each term, the LLMs either identified an equivalent Banjarese word or flagged it as a lexical gap if no direct translation was available.

In the second experiment, *Banjarese-to-Indonesian*, we selected fifty Banjarese-specific terms from the Banjarese dataset. These terms, along with the complete Indonesian dataset, were processed using GPT-4, DeepSeek, and Gemini. Each model annotated the Banjarese terms by either providing equivalent Indonesian translations or identifying lexical gaps.

The annotations were validated by two native speakers: a university instructor specializing in linguistics, who is a native Banjarese speaker, reviewed the Banjarese annotations, while an Indonesian master's student specializing in lexical semantics validated the Indonesian annotations. The accuracy results of each model's performance for both language pairs in the experiments are presented in Table 9.4, as evaluated by the validators. Similar to the English–Arabic case study, the model exhibited two common errors: it either provided *incorrect TL words* or produced *literal translations* instead of identifying lexical gaps.

For instance, the Indonesian word *beras*, meaning "*uncooked rice*", was incorrectly translated by GPT-4 into Banjarese as *karak*, which refers to "*hardened rice*". Similarly, DeepSeek mistranslated the Indonesian word *papaya*, meaning "*a large oval tropical fruit with yellowish flesh*", as *gandis*, which denotes "*yellow mangosteen*" in Banjarese. Likewise, Gemini erroneously translated *balinjan* meaning "*tomato*" as *terung* "*eggplant*". Additionally, GPT-4 rendered the Banjarese term *janar* "*turmeric*" as *jahe* in Indonesian, failing to capture the nuanced distinction between the two terms.

These examples, along with those from the English-Arabic case study, highlight the limitations of the three LLMs in cross-linguistic lexical annotation, particularly in handling culturally specific or nuanced terms.

Part IV

Resulting Datasets and Applications

10 Resulting Datasets

Contents

This chapter presents the outcomes of dataset collection and development efforts conducted using the expert-sourcing and crowdsourcing methodologies. These resulting datasets serve as critical resources for linguistic research, computational lexicon development, and cross-linguistic studies. By analyzing the lexical diversity and semantic relationships, the chapter provides a detailed overview of the constructed datasets and conceptual overlaps across examined languages.

This chapter is divided into four main sections. Section 10.1 examines datasets constructed through expert-sourcing studies, highlighting the linguistic diversity of kinship terms across Arabic dialects and Indonesian languages, along with their enrichment in multilingual lexicons. Additionally, it presents datasets generated from experiments designed to identify basic-level categories and enhance resources like the Arabic WordNet.

Section 10.2 shifts the focus to datasets developed through crowdsourcing experiments. These include bilingual datasets across languages such as English, Arabic, Indonesian, Banjarese, Turkish, and Persian, with a particular emphasis on food and kinship terminology. It also includes monolingual datasets created by comparing kinship concepts with Turkish, Persian, and Maba terms.

Section 10.3 provides guidance on accessing these datasets, which are publicly available through the DataScientia platform. These datasets are designed to support ongoing and future research in computational linguistics, semantic analysis, and cross-cultural studies.

Finally, Section 10.4 explores the linguistic insights derived from these datasets. The analysis includes both expert-curated and crowdsourced datasets, revealing patterns of shared terminology and lexical variations across the

studied languages. This section highlights the linguistic richness of the datasets and illustrates their value in uncovering cultural and semantic nuances.

10.1 Constructed datasets from expert-sourcing studies

10.1.1 Lexical diversity in kinship across Arabic dialects

In this study, the overall contribution collection effort resulted in 180 words, 1,108 lexical gaps, and 19 new concepts identified, formalized, and collected in the Arabic dialects.

Detailed statistics about the collected gaps and words are shown in Table 10.1. New concepts were identified in three subdomains: siblings, cousins, and grandchildren. The total number of new concepts, 19, is lower than the sum of new concepts per language due to overlaps across languages: for example, أَخٌ في الرضاعة meaning "*breastfeeding brother*" was found in all seven dialects, أخت لأم meaning "*maternal sister*" was found both in Syrian and in Egyptian, while أَبْيِه meaning "*elder cousin, son of mother's brother*" only exists in Egyptian.

Dialects	Words	Gaps w/o new concepts	New concepts	Gaps considering new concepts
Algerian	28	156	8	165
Egyptian	32	152	10	152
Moroccan	22	162	8	169
Palestinian	23	161	10	166
Syrian	24	160	9	169
Tunisian	23	161	2	178
Gulf	28	156	10	169
Total	**180**	**1108**	**19 (unique)**	**1168**

Table 10.1: The count of the diversity items collected and identified in the Arabic dialects

For more details, refer to the dataset provided in Table 10.18. This dataset, available in the DataScientia repository, contains data generated from the experiment.

10.1.2 Lexical diversity in kinship across Indonesian languages

In this study, the overall contribution effort resulted in the collection of 41 words and the identification of 517 lexical gaps in Indonesian, Javanese, and Banjarese. Additionally, three new, yet unattested word meanings were also discovered and formalised as new concepts. All three are used in Banjarese in the uncle/aunt subdomain:

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

Statistics on the data collected for each language are shown in Table 10.2.

For more details, refer to the dataset provided in Table 10.18. This dataset, available in the DataScientia repository, contains data generated from the experiment.

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

Table 10.2: The count of the diversity items collected and identified in the Indonesian languages

10.1.3 Enrichment of KinDiv: a multilingual lexicon for kinship

Our contribution to the development of the KinDiv lexicon included 396 words and 1,503 lexical gaps, provided by the native speakers of the ten languages: Arabic, English, French, Hindi, Hungarian, Italian, Kannada, Malayalam, Mongolian, and Spanish. Detailed statistics on the collected words and gaps are presented in Table 10.3.

Moreover, 107 new words and collocations were collected. Notably, many missing terms were expressed through restricted collocations in Spanish, Mongolian, and Hungarian. For example, Mongolian speakers contributed 23 new inputs, including restricted collocations such as ач хүү "*son's son*" and нагац ах "*maternal uncle*". Similarly, Spanish examples include *hermano menor* "*younger brother*" and *tío materno* "*maternal uncle*", while

Language	Words	Gaps
Arabic	28	162
English	32	154
French	36	158
Hindi	45	140
Hungarian	35	155
Italian	26	160
Kannada	64	128
Malayalam	63	131
Mongolian	35	151
Spanish	32	164
Total	396	1503

Table 10.3: Statistics of the collection of diversity data by language

Hungarian contributions included *fiúunoka* "*grandson*" and *nagytestvér* "*elder sibling*". Additionally, French contributors identified certain colloquial terms and morphological alternations, such as *tata* "*aunt*" and *papi* "*grandfather*", as well as *aîné* "*elder brother*".

For further details on the gathered contributions, refer to the dataset uploaded and integrated with the KinDiv resource[1], which is also available for direct download[2]. Additionally, this dataset can be accessed through the experiment's page on DataScientia, as presented in Table 10.18.

10.1.4 Experiment: basic-level categories

In this study, the overall contribution effort resulted in the collection of 5,649 words and the identification of 159 lexical gaps across Arabic, Turkish, Persian, Indonesian, Banjarese, and Javanese. A summary of the data collected for each language is shown in Table 10.4.

For more details, refer to the dataset provided in Table 10.18. This dataset includes six spreadsheets—one for each language examined in this experiment—and is available in the DataScientia repository. It contains data generated from the experiment.

10.1.5 Advancing the Arabic WordNet: elevating content quality

The overall effort to collect contributions resulted in updating 5,554 synsets from AWN V1. We added 2,726 new lemmas, 9,322 new glosses, and

[1] http://ukc.disi.unitn.it/index.php/kinship/
[2] https://github.com/kbatsuren/KinDiv

Language	Gaps	Words
Arabic	28	940
Turkish	15	953
Persian	65	903
Indonesian	20	948
Javanese	19	949
Banjarese	12	956
Total	159	5,649

Table 10.4: Summary of the collection of gaps and terms by language

12,204 new example sentences. We also identified 236 lexical gaps and inserted 701 phrasets. Furthermore, we deleted 8751 incorrect lemmas. More details regarding the counts of these contributions are presented in Table 10.5.

For more details, refer to the dataset provided in Table 10.18. This dataset, available in the DataScientia repository, contains data generated from the experiment. For each POS, two spreadsheets were uploaded to DataScientia; the first file includes the final resulting Arabic synsets, and the second contains the added and deleted synset components.

	Noun	Verb	Adj	Adv	Total
Updated synsets	3,938	1,364	181	71	**5,554**
New lemmas	2,581	64	72	9	**2,726**
Deleted lemmas	6,050	2,387	223	91	**8,751**
New glosses	6,511	2,258	446	107	**9,322**
New examples	7,597	3,620	782	205	**12,204**
Gaps	28	187	0	21	**236**
Phrasets	364	275	0	62	**701**

Table 10.5: Statistics of the data addition and deletion into/from AWN

10.2 Crowdsourced datasets

10.2.1 English vs Arabic: food

This section presents the datasets generated from two crowdsourcing experiments conducted on food-related terms between English and Arabic. The first experiment focused on the crowdsourcing diverse data from English to Arabic, where the identified gaps and equivalent word meanings were collected in Arabic. The second experiment proceeded in the reverse direction,

with the gathered information recorded in English.

English → Arabic

This experiment contributed to the identification of 1,532 lexical gaps, 832 equivalent Arabic words, and 100 new Arabic words (absent from the original Arabic input dataset but present in the supra-lingual layer of the UKC), achieving an average Alpha score of 0.84. More information on the crowdsourced Arabic data is provided in Table 10.6 and Table 10.7.

For more details, refer to the dataset provided in Table 10.17, where English represents the SL and Arabic represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

	Gaps		**Words**		**New Concepts**		**Alpha**	
Task	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		
Average							0.86	0.81

Table 10.6: Information on the crowdsourced data by G_1 and G_2 in Arabic.

Arabic → English

This experiment identified 608 lexical gaps, 298 equivalent English words, and 49 new food concepts (absent from both the original English input dataset and the supra-lingual layer of the UKC), achieving an average Alpha score of 0.85. Examples of these new concepts include: خشاف, meaning "*a drink made from raisins, figs, or similar dried fruits, sweetened and soaked or boiled in water, typically consumed by Muslims during Ramadan*"; الأسمران, meaning "*water and wheat*"; الأحمران, meaning "*meat and bread*"; and الأبيضان, meaning "*water and yogurt*". The exploration of such concepts is one of the primary aims of this experiment, contributing to the enrichment of the supra-lingual layer of the UKC and enhancing the accuracy of the concept hierarchy within this layer by incorporating these concepts into its restructuring. More information on the crowdsourced English data is provided in Table 10.8.

	Gaps		Words		New Concepts		Alpha	
Task	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

Table 10.7: Information on the crowdsourced data by G_3 and G_4 in Arabic.

For more details, refer to the dataset provided in Table 10.17, where Arabic represents the SL and English represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

10.2.2 Indonesian vs Banjarese: food

This section presents the datasets generated from two crowdsourcing experiments on food-related terms between Indonesian and Banjarese. The first

	Gaps				Words				New Concepts				Alpha			
Task	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
1	29	19	21	26	6	16	14	9	3	2	0	2	0.87	0.92	0.81	0.91
2	25	24	18	26	10	11	17	9	1	0	2	3	0.75	0.92	0.89	0.95
3	24	19	32	22	11	16	3	13	2	1	3	1	0.82	0.85	0.72	0.88
4	27	28	15	24	8	7	20	11	2	2	3	2	0.84	0.80	0.88	0.87
5	23	25	21	22	12	10	14	13	1	2	4	2	0.89	0.81	0.87	0.85
6	27	19	15	27	8	16	20	8	2	2	2	1	0.83	0.77	0.92	0.81
7	21	29	-	-	14	2	-	-	1	3	-	-	0.77	0.85	-	-
Total	176	163	122	147	69	78	88	63	12	12	14	11				
Average													0.82	0.85	0.85	0.88

Table 10.8: Information on the crowdsourced data in English.

experiment focused on collecting diverse data by crowdsourcing from Indonesian to Banjarese, where the identified gaps and equivalent word meanings were collected in Banjarese. The second experiment proceeded in the reverse direction, with the gathered information recorded in Indonesian.

Indonesian → Banjarese

This experiment identified 750 lexical gaps, 507 equivalent words, and 43 new Banjarese words (missing from the original Banjarese input dataset), with an average Alpha score of 0.83. Examples of these new words include: *alai*, meaning "*a fruit that is a long, flattened legume pod up to 36 centimeters long, containing up to 21 hard, black seeds, each around 2 centimeters in length*"; and *dadih*, meaning "*water, cow's milk, buffalo milk, etc., that has been concentrated or thickened*". More information on the crowdsourced Banjarese data is provided in Table 10.9.

For more details, refer to the dataset provided in Table 10.17, where Indonesian represents the SL and Banjarese represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

	Gaps		**Words**		**New Concepts**		**Alpha**	
Task	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 10.9: Information on the crowdsourced data in Banjarese.

Banjarese → Indonesian

This experiment identified 201 lexical gaps, 98 equivalent words, and 30 new Indonesian words (missing from the original Indonesian input dataset), achieving an average Alpha score of 0.83. Examples of these new words include: *palajau*, meaning "*a type of fruit that grows by the river and is used as an ingredient in vegetable dishes*"; and *wajik tuba*, meaning "*a type of pastry made from glutinous rice and brown sugar*". More information on the crowdsourced Indonesian data is presented in Table 10.10.

For more details, refer to the dataset provided in Table 10.17, where Banjarese represents the SL and Indonesian represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

	Gaps		**Words**		**New Concepts**		**Alpha**	
Task	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		
Average							0.83	0.83

Table 10.10: Information on the crowdsourced data in Indonesian.

10.2.3 English vs. Persian: kinship

As introduced in Section 9.3, two crowdsourcing experiments were conducted to explore English and Persian kinship terms. The first experiment involved comparing English to Persian, where the identified gaps and words were in Persian. The second experiment focused on the reverse direction, comparing Persian to English, with the identified gaps and words in English. The contributions collected through these experiments are presented in Table 10.11.

For further details, refer to the two datasets: the first is the Persian dataset, and the second is the English dataset. Their links are provided in Table 10.17. The datasets, available in the DataScientia repository, contain the data generated from this experiment.

10.2.4 Arabic vs. Persian: kinship

As described in Section 9.4, two crowdsourcing experiments were conducted to explore Arabic and Persian kinship terms. The first experiment involved

	Gaps	Words	New Words	Alpha
English → Persian	9	8	0	1.0
Persian → English	27	3	3	0.89
Total	**36**	**11**	**3**	**0.95 (avg)**

Table 10.11: Information on the crowdsourced kinship data in English and Persian.

comparing Arabic to Persian, where the identified gaps and words were in Persian. The second experiment focused on the reverse direction, comparing Persian to Arabic, with the identified gaps and words in Arabic. The contributions collected through these experiments are presented in Table 10.12.

For further details, refer to the two datasets: the first is the Persian dataset, and the second is the Arabic dataset. Their links are provided in Table 10.17. The datasets, available in the DataScientia repository, contain the data generated from this experiment.

	Gaps	Words	New Words	Alpha
Arabic → Persian	16	22	0	0.79
Persian → Arabic	13	3	3	1.0
Total	**29**	**25**	**3**	**0.90 (avg)**

Table 10.12: Information on the crowdsourced kinship data in Arabic and Persian.

10.2.5 English vs. Turkish: kinship

As introduced in Section 9.5, two crowdsourcing experiments were performed to investigate English and Turkish kinship terms. The first experiment involved comparing English to Turkish, where the identified gaps and words were in Turkish. The second experiment focused on the reverse direction, comparing Turkish to English, with the identified gaps and words in English. The contributions collected through these experiments are presented in Table 10.13.

For further details, refer to the two datasets: the first is the Turkish dataset, and the second is the English dataset. Their links are provided in Table 10.17. The datasets, available in the DataScientia repository, contain the data generated from this experiment.

	Gaps	Words	New Words	Alpha
English → Turkish	6	11	0	0.94
Turkish → English	24	3	1	0.90
Total	**30**	**14**	**1**	**0.92 (avg)**

Table 10.13: Information on the crowdsourced kinship data in English and Turkish.

10.2.6 Kinship concepts vs. Turkish

The overall results of this experiment include the identification of 164 lexical gaps, 33 equivalent Turkish words, and four new kinship concepts in Turkish. These concepts are: ***ata***, meaning "*any male ancestor such as father, grandfather, or great-grandfather*"; ***Üvey kardeş*** meaning "*half-sibling (one parent shared)*"; ***Üvey abla*** meaning "*elder female step-sibling (step elder sister)*"; and ***Üvey abi, üvey ağabey***, meaning "*elder male step-sibling (step elder brother)*". Table 10.14 provides more information on the crowdsourced contributions by subdomain in Turkish.

	Gaps	Words	New Concepts	Alpha
grandparents	15	4	1	0.83
grandchildren	25	3	0	0.84
siblings	24	5	3	0.87
uncle/aunt	23	4	0	0.84
nephew/niece	34	3	0	0.90
cousins	43	14	0	0.95
Total	**164**	**33**	**4**	**0.88 (avg)**

Table 10.14: Information on the crowdsourced data for kinship concepts in Turkish.

For more details, refer to the dataset provided in Table 10.17, where "*Concepts*" represents the SL and Turkish represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

10.2.7 Kinship concepts vs. Persian

The overall collected contributions of this experiment contain the identification of 163 lexical gaps, 34 equivalent words in Persian, and no new kinship concepts. Table 10.15 provides more information on the crowdsourced contributions by subdomain in Persian.

For more details, refer to the dataset provided in Table 10.17, where "*Concepts*" represents the SL and Persian represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

	Gaps	**Words**	**Alpha**
grandparents	13	6	0.93
grandchildren	24	4	0.77
siblings	23	6	0.82
uncle/aunt	23	4	1.00
nephew/niece	35	2	0.80
cousins	45	12	0.85
Total	**163**	**34**	**0.85 (avg)**

Table 10.15: Information on the crowdsourced data for kinship concepts in Persian.

10.2.8 Kinship concepts vs. Maba

The overall collected contributions of this experiment contain the identification of 177 lexical gaps, 20 equivalent words in Maba, and no new kinship concepts. Table 10.16 provides more information on the crowdsourced contributions by subdomain in Maba.

For more details, refer to the dataset provided in Table 10.17, where "*Concepts*" represents the SL and Maba represents the TL. This dataset, available in the DataScientia repository, contains data generated from the experiment.

	Gaps	**Words**	**Alpha**
grandparents	17	2	1.00
grandchildren	27	1	1.00
siblings	27	2	1.00
uncle/aunt	22	5	0.78
nephew/niece	33	4	0.83
cousins	51	6	0.81
Total	**177**	**20**	**0.90 (avg)**

Table 10.16: Information on the crowdsourced data for kinship concepts in Maba.

10.3 Downloading datasets from DataScientia

As introduced in Section 8.4, two websites integrated with DataScientia provide comprehensive details about our experiments. The first is the *LingoGap website*, which hosts the crowdsourced datasets generated through crowdsourcing experiments. The second is the *Expert-sourcing lexical diversity project*, where we publish datasets constructed during expert-sourcing studies. Figure 10.1 presents an excerpt from the bilingual experiments conducted using LingoGap.

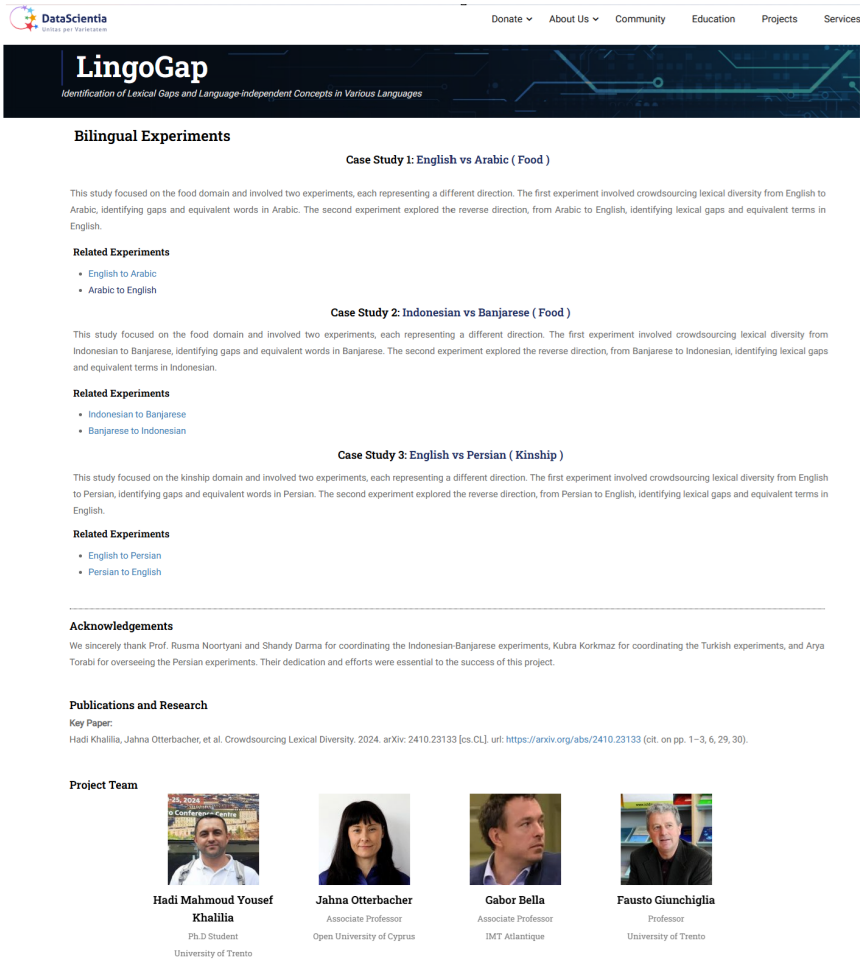


Figure 10.1: A screenshot from the LingoGap page showing some bilingual experiments conducted using LingoGap

As shown in Figure 10.1, on the LingoGap website, each case study includes two experiments conducted between language pairs, with each experiment representing one direction—from the SL to the TL. For example, the first case study, English vs. Arabic (Food), contains two related exper-

iments: (1) English to Arabic, where lexical gaps and their equivalents are explored in Arabic, and (2) Arabic to English, where gaps and words are identified in English.

For each experiment, the LingoGap website provides detailed information, including the experiment title, project status, description, start date, end date, experiment goals and tasks, eligibility criteria, and a link to the resulting dataset. This is illustrated in Figure 10.2, which shows a screenshot of the English to Arabic experiment page on LingoGap, including a link to the resulting crowdsourced Arabic dataset.

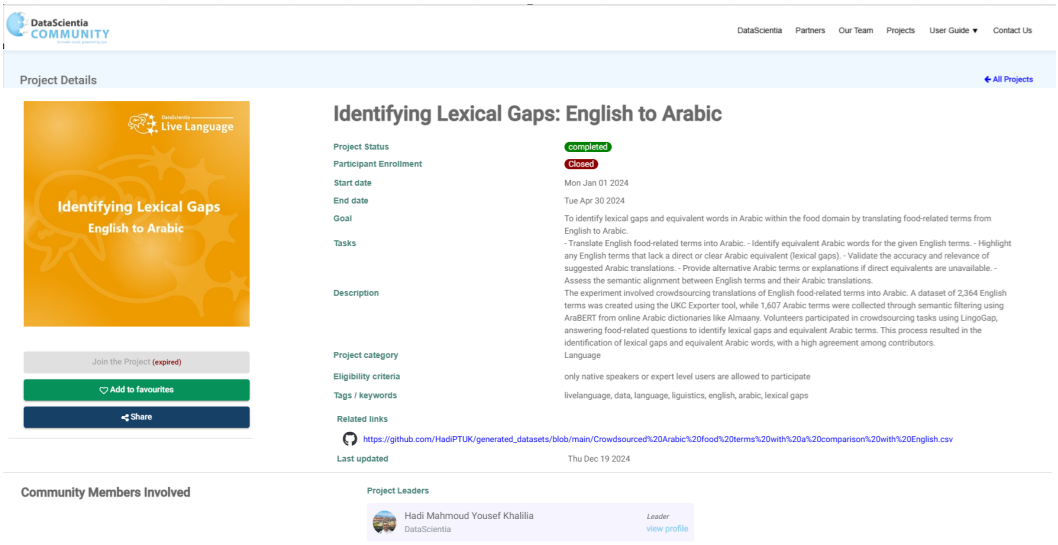


Figure 10.2: The English-to-Arabic experiment page on LingoGap, integrated with DataScientia

Table 10.17 presents the links to datasets generated from crowdsourcing experiments conducted using LingoGap. These links are made available on the LingoGap website via DataScientia.

Similarly, the Expert-sourcing lexical diversity website presents five case studies along with the related experiments conducted using the expert-driven methodology. For each experiment, the website offers information similar to that provided by the LingoGap website, including an experiment description, goals and tasks, and a link to the resulting dataset. This is illustrated in Figure 10.3, which shows a screenshot of the page dedicated to the experiment of exploring lexical diversity in kinship across Indonesian languages on the Expert-sourcing lexical diversity website, including a link to the collected dataset.

Table 10.18 presents links to datasets collected through the expert-sourcing studies. These links are accessible on the Expert-sourcing lexical diversity website via DataScientia.

Experiment	Source Language	Target Language	Link
English vs Arabic: Food	English	Arabic	https://ds.datascientia.eu/community/public/projects/aef86b5b-5742-4529-9c46-71b5142f91bd
English vs Arabic: Food	English	Arabic	https://ds.datascientia.eu/community/public/projects/db9ec453-4775-43d6-a89a-85a9600b9781
Indonesian vs Banjarese: Food	Indonesian	Banjarese	https://ds.datascientia.eu/community/public/projects/c3d242b7-4ddc-419f-bffd-38d9f9760a05
Indonesian vs Banjarese: Food	Banjarese	Indonesian	https://ds.datascientia.eu/community/public/projects/eacdb797-f8bf-45d4-8909-a4fd50ae8910
English vs. Persian: Kinship	English	Persian	https://ds.datascientia.eu/community/public/projects/bedff19e-40ec-42d0-9600-1cd1014cf09b
English vs. Persian: Kinship	Persian	English	https://ds.datascientia.eu/community/public/projects/34e04eb6-c776-4e54-a0e6-465a72f6031f
Arabic vs. Persian: Kinship	Arabic	Persian	https://ds.datascientia.eu/community/public/projects/65467d7d-35fa-4756-a840-83c1eb6abad6
Arabic vs. Persian: Kinship	Persian	Arabic	https://ds.datascientia.eu/community/public/projects/44cb6bef-54f0-47cd-80b5-86e25a86f4ab
English vs. Turkish: Kinship	English	Turkish	https://ds.datascientia.eu/community/public/projects/c9d0a079-56de-4878-a164-ebe75c61910f
English vs. Turkish: Kinship	Turkish	English	https://ds.datascientia.eu/community/public/projects/74f2f22a-ef58-4a8a-ae19-6aa7e55c5da0
Kinship concepts vs. Turkish	Concepts	Turkish	https://ds.datascientia.eu/community/public/projects/95718721-14dd-4c6e-a3c3-ab7a3e2cf4bc
Kinship concepts vs. Persian	Concepts	Persian	https://ds.datascientia.eu/community/public/projects/6ca5612a-1bd8-4b1d-b69e-2a4e32ce3521
Kinship concepts vs. Maba	Concepts	Maba	https://ds.datascientia.eu/community/public/projects/c0fce09e-86e0-4d28-80eb-a823097d7a37

Table 10.17: Links to datasets crowdsourced using LingoGap, offered through DataScientia

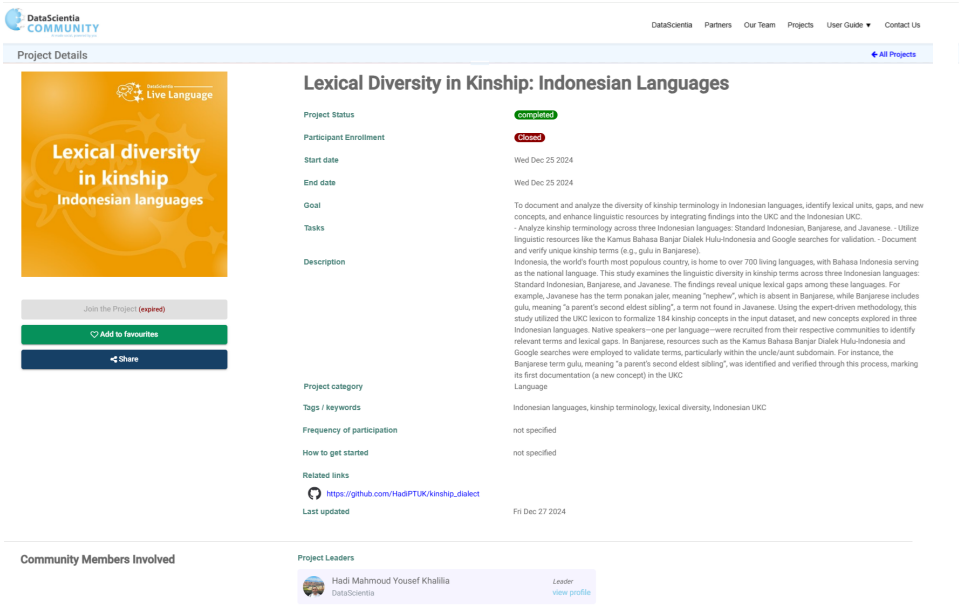


Figure 10.3: The page for the "Exploring Lexical Diversity Across Indonesian Languages" experiment on the Expert-Sourcing Lexical Diversity website, integrated with DataScientia

Experiment	Link
Lexical diversity in kinship across Arabic dialects	https://ds.datascientia.eu/community/public/projects/f402520c-aaf2-40cc-9736-93a1b1ff02c8
Lexical diversity in kinship across Indonesian languages	https://ds.datascientia.eu/community/public/projects/a131c694-4b75-409d-937c-f1560c2c7de5
Enrichment of KinDiv: A multilingual lexicon for kinship	https://ds.datascientia.eu/community/public/projects/130c9396-c939-406d-bb7b-79f9b7d4ae3c
Lexical diversity in the basic-level category across six languages	https://ds.datascientia.eu/community/public/projects/4ea0a2de-caf0-4d82-b27b-7243f4146f9f
Advancing the Arabic wordnet: elevating content quality	https://ds.datascientia.eu/community/public/projects/7f152487-df1b-409d-830a-73b531e2cee3

Table 10.18: Links to datasets collected through expert-sourcing studies, offered through DataScientia

10.4 Linguistic insights

This section examines key findings on how languages convey ideas, with a focus on the relationships within and between languages. By analyzing

examples from the produced diversity-aware datasets—both expert-curated and crowdsourced—in the semantic domains of kinship and food terminology, we uncover shared linguistic meanings across various dialects and languages. Specifically, we explore four cases: (1) shared kinship terminology across seven Arabic dialects, (2) shared kinship terminology across three Indonesian languages, (3) overlapping food terminology between English and Arabic, and (4) overlapping food terminology between Indonesian and Banjarese.

10.4.1 Expert-curated datasets

Shared kinship terminology across Arabic dialects

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

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

$$\text{overlap}(d, l_a, ..., l_n) = \frac{|\text{LexCons}(d, l_a) \cap ... \cap \text{LexCons}(d, l_n)|}{\max(|\text{LexCons}(d, l_a)|, ..., |\text{LexCons}(d, l_n)|)} \quad (10.1)$$

Figure 10.4 shows the overlaps between pairs of Arabic dialects over the kinship domain. For example, the intersection of Egyptian and Gulf dialects gives a shared coverage of 74.5%, while all dialects are 47.1% the same. In the former case, the number of lexicalisations in Egyptian is 51, and in Gulf is 42. Also, 38 of these lexical units are included in both dialects; see the dataset uploaded to GitHub[3]. For example, Equation (10.1) calculates the overlap between Egyptian and Gulf (both represented in ISO 639-3[4] as "arz" and "afb", respectively) in the Kinship domain (K) as follows:

[3] https://github.com/HadiPTUK/kinship_dialect
[4] https://iso639-3.sil.org

$$\begin{aligned}\text{overlap}(K, \text{arz}, \text{afb}) &= \frac{|\text{LexCons}(K, \text{arz}) \cap \text{LexCons}(K, \text{afb})|}{\max(|\text{LexCons}(K, \text{arz})|, |\text{LexCons}(K, \text{afb})|)} \\ &= \frac{38}{\max(51, 42)} = \frac{38}{51} = 74.5\%\end{aligned} \tag{10.2}$$

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

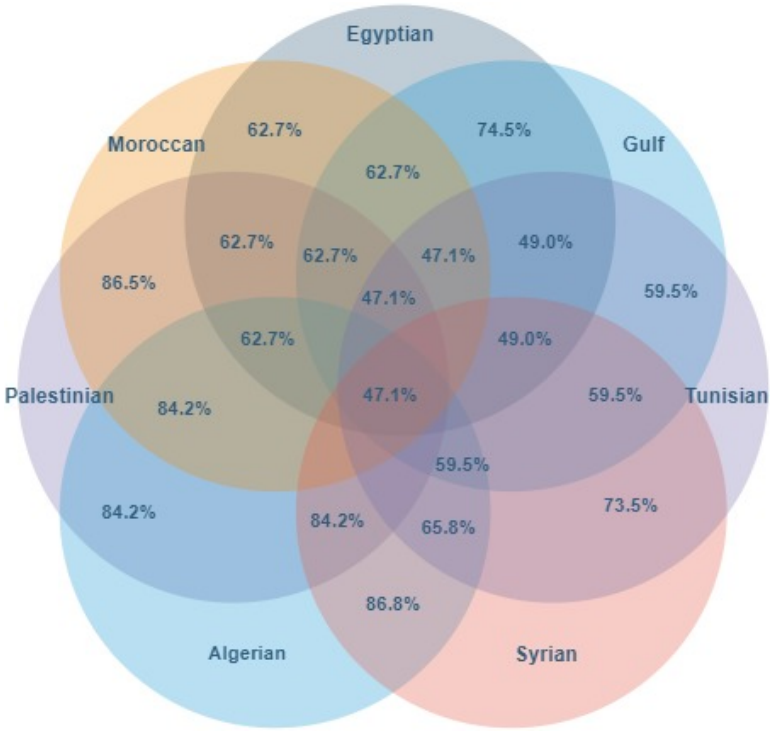


Figure 10.4: The overlap (percentage of shared lexicalisations) for Arabic dialects

We find these overlaps—e.g. an overlap of 59.5% between Gulf and Tunisian, or the overall overlap of 47.1% among all seven dialects—lower than our initial expectations on dialectal variations. Arab dialectologists justify such differences with two major factors: linguistic and religious influence [165]. By linguistic influence, we refer to the historical interaction of language-speaker communities, which affects the lexicons. Examples are the Egyptian dialect influenced by the Coptic language (historically spoken by the Copts, starting from the third century AD in Roman Egypt) or the Levantine dialect influenced by the Western Aramaic, Canaanite, Turkish, and Greek languages. The Gulf dialect is one of the Peninsular groups, which was influenced by South Arabian Languages. Secondly, the religion of

the speaker community also affects the lexicon. Religion is a sociolinguistic variable that shapes how Arabic is spoken. Religion in Arab countries is a matter of group affiliation and is not usually considered an individual choice: one is born a Muslim, Christian, Jew, or Druze, and this becomes a bit like one's ethnicity. So, for example, within the Egyptian speech community, one can find language mixing between Islamic and Christian terms, and the same in the Levantine community, which consists of a mixing of Muslims, Christians, Jews, and Druze. The Gulf communities, instead, mostly consist of Muslims [158].

Shared kinship terminology across Indonesian languages

More than 90% of our 184 initial kinship concepts were found to be gaps in the three Indonesian languages, as shown in Table 10.2. Using Equation (10.1), we calculated the overlaps between the Indonesian languages in terms of kinship lexicalisations, shown in Figure 10.5. For more details, see the dataset uploaded to the GitHub repository[5]. 35.3% of the concepts are lexicalised by the three Indonesian languages studied. The Javanese–Banjarese overlap is 52.9%, Javanese–Indonesian is 60%, and finally Banjarese–Indonesian is 41.2%. Even though all three languages are included in the Malayo-Polynesian branch of the Austronesian language family, Indonesian and Banjarese are considered Malayic languages, while Javanese is not, which is the first reason for this result. Furthermore, these languages exist on different islands in Indonesia; Javanese exists on Java Island, Banjarese is located on the southern part of Borneo Island, and the Indonesian language is based on Malay, which is spoken on Sumatra Island [139], so this geographical barrier restricts interactions between speakers, and each language has developed within its own speech community.

Finally, using Equation (10.1), we computed the overlaps between Arabic dialects and Indonesian languages. Figure 10.6 shows that the ten languages together cover only 3.9% of the concepts, and the most similar language pair, namely Egyptian–Indonesian, is merely 5.9% similar. For researchers in ethnography or comparative linguistics, the observation of such pronounced levels of cross-lingual and cross-cultural diversity may not come as a surprise, as major variations in kin patterns are well known in these domains. On the other hand, we believe that beyond these narrow fields of research, there is a general lack in understanding the depth of diversity in how, through languages, people describe and interpret the world. Most computational linguists and engineers who build language processing systems, as well as the users who trust such systems for their daily activities, do not suspect the breadth of the mental divide across languages that language applications, such as machine translation systems, are meant to bridge. We think that

[5] https://github.com/HadiPTUK/kinship_dialect

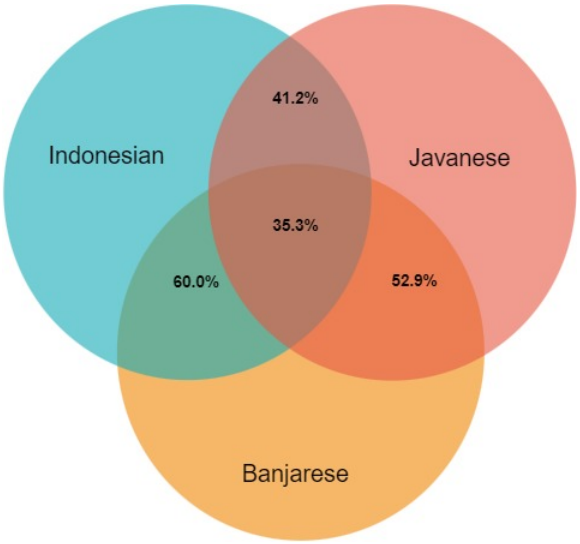


Figure 10.5: The number of words in the intersection of Indonesian languages according to shared meaning

through quantified measures, as we are attempting to do with our simple measure of overlap introduced on p. 152, can be useful to improve our qualitative grasp on diversity, which we consider a promising direction for future research.

Table 10.19 includes statistics of collected words and gaps by domain across Arabic and Indonesian languages. The results show that only three words in the domain of cousins are identified in the Indonesian languages, while in Egyptian, 16 words are used around the concept of the cousin.

Domains	**Words**	**Gaps**
Grandparents	21	169
Grandchildren	19	251
Siblings	37	173
Uncle/Aunt	44	226
Nephew/Niece	33	297
Cousins	67	503
Total	221	1619

Table 10.19: Statistics of the collection of diversity data by domain

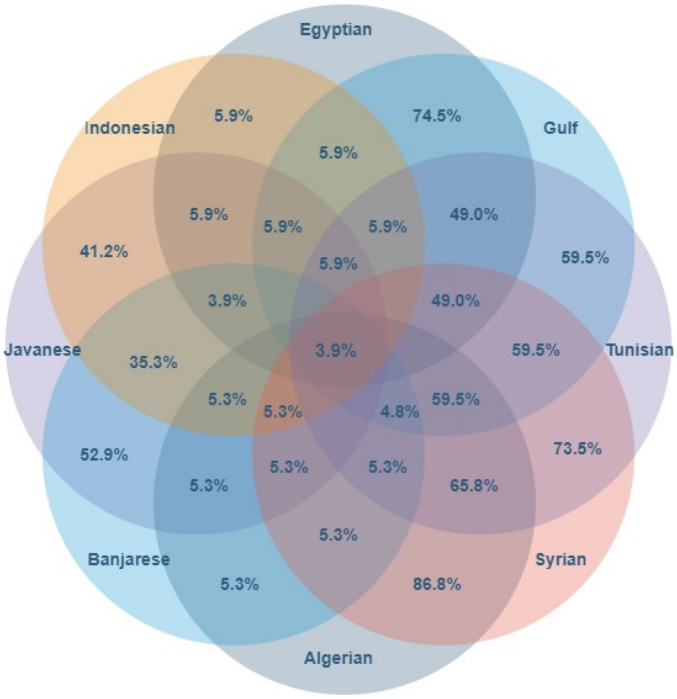


Figure 10.6: The number of words in the intersection of Indonesian and Arabic languages according to shared meaning

10.4.2 Crowdsourced datasets

Exploring English-Arabic food term overlaps

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.

$$\text{overlap}(d, l_A, l_B) = \frac{|\text{LexCons}(d, l_A) \cap \text{LexCons}(d, l_B)|}{\max(|\text{LexCons}(d, l_A)|, |\text{LexCons}(d, l_B)|)} \tag{10.3}$$

Figure 10.7 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 2413 (2364 words in the English input dataset and 49 new words, which are missing from the English input dataset), and in Arabic is 1707 (1607 words in the Arabic input dataset and 100 new words). Also, 1130 (832 words were explored in the experiment of (English to Arabic), 298 words were identified in the experiment of (Arabic to English) of these lexical units are included

in both languages. For example, Equation (10.3) 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:

$$\text{overlap}(F, \text{eng}, \text{arb}) = \frac{|\text{LexCons}(F, \text{eng}) \cap \text{LexCons}(F, \text{arb})|}{\max(|\text{LexCons}(F, \text{eng})|, |\text{LexCons}(F, \text{arb})|)} \quad (10.4)$$

$$\text{overlap}(F, \text{eng}, \text{arb}) = \frac{1130}{\max(2413, 1707)} = \frac{1130}{2413} = 46.8\% \quad (10.5)$$

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 [4, 8]. By linguistic influence, we refer to the etymological origin and borrowing of the language, which 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 [8].

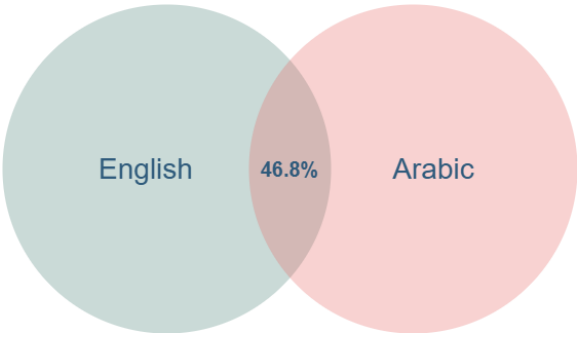


Figure 10.7: The overlap (percentage of shared lexicalizations) for English and Arabic languages

Exploring Indonesian-Banjarese food term overlaps

Figure 10.8 shows the overlaps between Indonesian and Banjarese over the food domain, as shown, the intersection of Indonesian and Banjarese languages gives a shared coverage of 40.9%. The number of lexicalisations in

Indonesian is 1,478 (1,448 words in the Indonesian input dataset and 30 new words, which are missing from the Indonesian input dataset), and in Banjarese is 855 (812 words in the Banjarese input dataset and 43 new words). Also, 605 (507 words were explored in the experiment of (Indonesian to Banjarese), 98 words were identified in the experiment of (Banjarese to Indonesian) of these lexical units are included in both languages. For example, Equation (10.3) 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:

$$\text{overlap}(F, \text{ind}, \text{bjn}) = \frac{|\text{LexCons}(F, \text{ind}) \cap \text{LexCons}(F, \text{bjn})|}{\max(|\text{LexCons}(F, \text{ind})|, |\text{LexCons}(F, \text{bjn})|)} \tag{10.6}$$

$$\text{overlap}(F, \text{ind}, \text{bjn}) = \frac{605}{\max(1478, 855)} = \frac{605}{1478} = 40.9\% \tag{10.7}$$

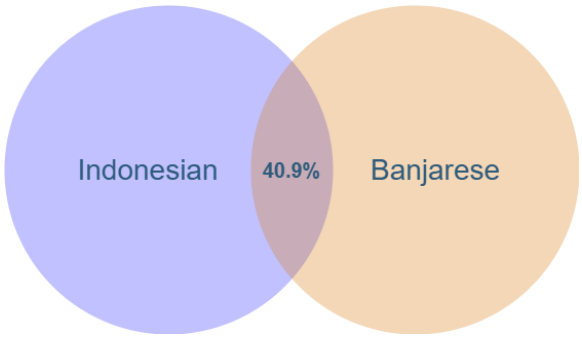


Figure 10.8: The overlap (percentage of shared lexicalizations) for Indonesian and Banjarese languages.

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 [139], 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.

11

Applications of Diversity-Aware LSRs

Contents

This chapter explores key implementations of diversity-aware lexical-semantic resources, focusing on their integration and utilization across different languages and platforms. It begins with an overview of UKC data integration in Section 11.1, highlighting how the UKC serves as a foundational framework for integrating the diversity-aware datasets generated from our experiments. This is followed by Section 11.2, which introduces the Arabic UKC—an instance of the UKC specifically developed for Arabic dialects. The section describes the integration of kinship contributions collected from Arabic dialects in the experiment described in Section 5.1 with the Arabic UKC.

Section 11.3 presents the Indonesian UKC—another UKC instance developed for Indonesian languages. It describes the integration of Indonesian kinship contributions collected in the experiment introduced in Section 5.2 with the Indonesian UKC.

In Section 11.4, a new version of the Arabic WordNet (AWN) is introduced—the first diversity-aware WordNet designed for the Arabic language. Finally, the chapter delves into applications in machine translation, demonstrating how diversity-aware lexical resources can improve accuracy and evaluation in machine translation systems, as discussed in Section 11.5.

11.1 UKC data integration

In this thesis, we utilize the UKC for various tasks. For example, we used it to create input datasets derived from its layers. Specifically, we extracted

kinship concepts from the super-lingual layer to serve as input datasets for experiments exploring linguistic diversity in kinship across Arabic dialects, Indonesian languages, Turkish, Persian, and Maba as described in Sections 5.1 to 5.3 and 9.6 to 9.8. Similarly, from the same layer, we constructed an input dataset of basic-level categories for an experiment investigating diversity in such concepts across Arabic, Indonesian, Javanese, Banjarese, Turkish, and Persian, as outlined in Section 5.4.

Additionally, we employed the lexicon layer to build datasets for specific languages. For instance, from the English lexicon, we extracted English food terms for comparison with Arabic food terms in the crowdsourcing experiment discussed in Section 9.1. Likewise, we extracted English kinship terms for comparisons with Persian and Turkish kinship lists, as detailed in Sections 9.3 and 9.5. We also retrieved English synsets for use in the experiment aimed at enhancing AWN quality, as explained in Section 5.5.

Furthermore, we leveraged the UKC to model the interlingual conceptual space, a critical component for systematically inferring lexical gaps. For this purpose, we adopted its kinship and food hierarchies in experiments related to kinship and food terminology.

We utilized the UKC not only as a foundational resource for dataset creation but also as a dynamic platform for integrating and storing the datasets resulting from our experiments. Its structured layers facilitated seamless information lookup, enabling efficient retrieval and comparison of linguistic data. Below, we provide examples that show the use of these features and services.

For instance, the 22 new kinship concepts identified during our experiments with Arabic dialects and Indonesian languages, as detailed in Sections 5.1 and 5.2, were integrated into the UKC by restructuring the domain hierarchy at the supra-lingual concept layer. The hierarchy of siblings, for example, was redesigned to include five new brotherhood concepts and five new sisterhood concepts. Specifically, in the Arabic-Egyptian lexicon, as shown in Figure 11.1, the term أَخٌ في الرضاعة, meaning "*breastfeeding brother*", was added as a sub-node under a newly created concept of brother, defined as "*a male person who has the same father, mother, or both parents as another person or has the same breastfeeding woman*". Also, from the figure, can be seen the terms أخ لأب meaning "*paternal brother*" and أخ لأم meaning "*maternal brother*" were inserted and connected to the half-brother concept. New concepts and lexicalizations are marked with white nodes and linked with blue lines.

The resulting lexical units and gaps were added into the UKC lexicons. The UKC website offers various services to system users, including browsable online access to database contents, source materials, and data visualization tools. The interactive exploration of linguistic diversity data in lexicons is the key feature of the website. Users can browse: (1) all meanings of a word

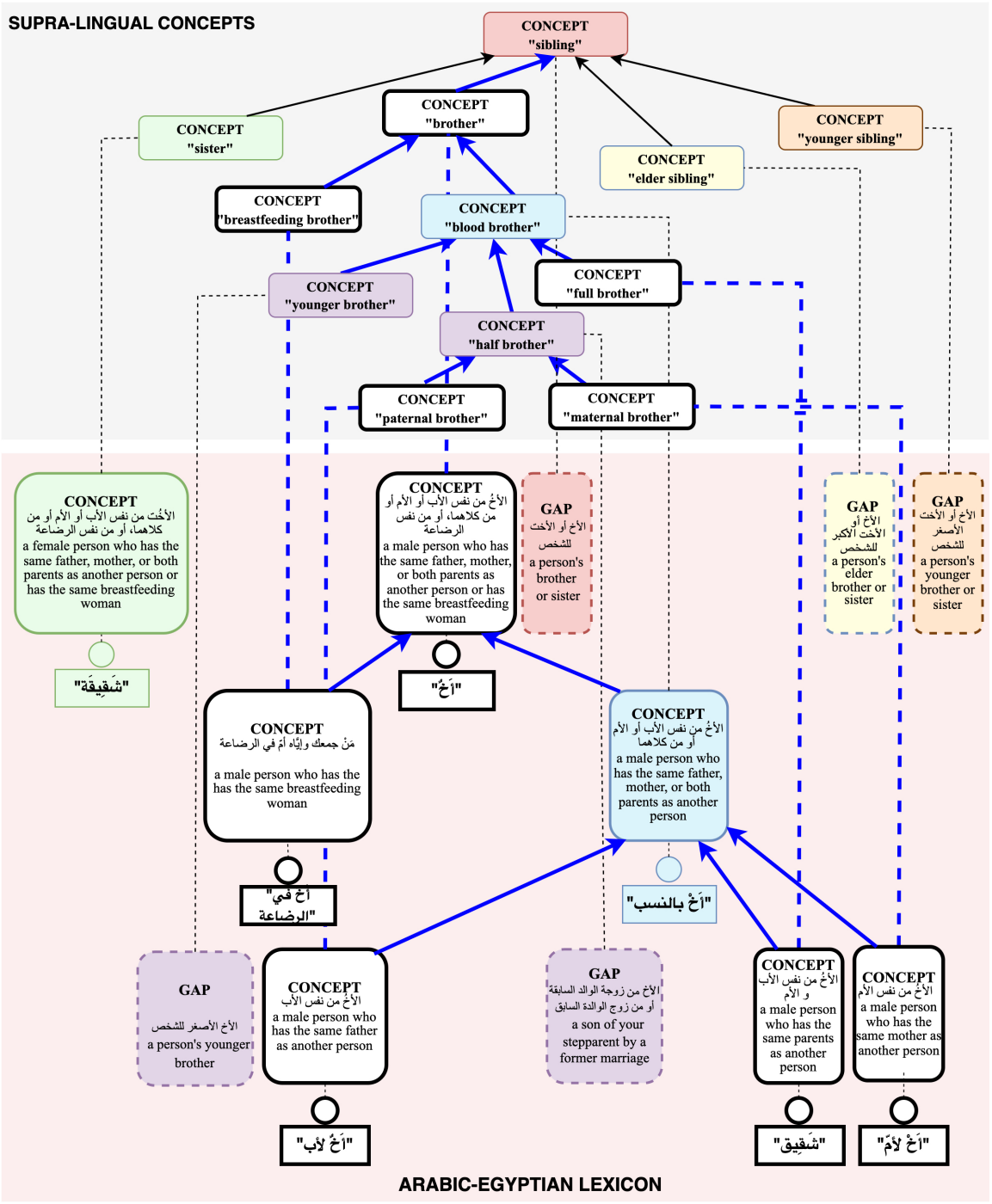


Figure 11.1: Structural elements in the UKC database after merging new concepts

within a selected language; and (2) lexicalizations and gaps of a concept across all languages contained in the database.

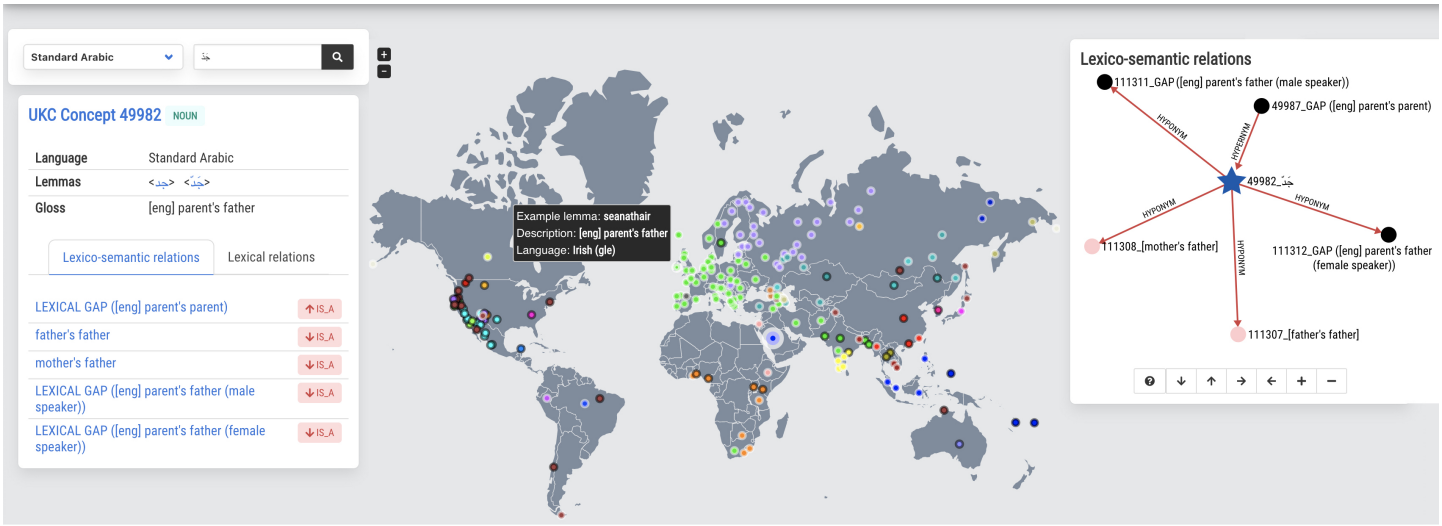


Figure 11.2: Exploring the concept of جَدّ as lexicalized in the Arabic language (left), in the world (middle), and as part of the shared concept hierarchy (right)

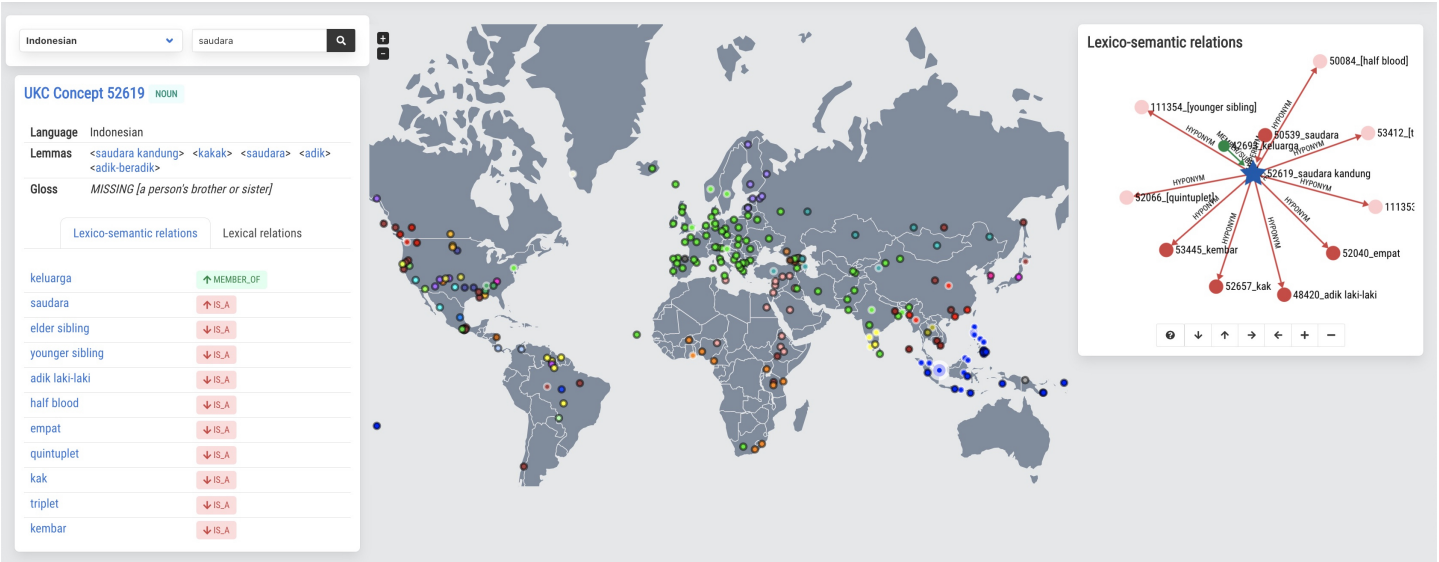


Figure 11.3: Exploring the concept of *saudara* as lexicalized in the Indonesian language (left), in the world (middle), and as part of the shared concept hierarchy (right)

Figure 11.2 shows a screenshot of the concept exploration functionality

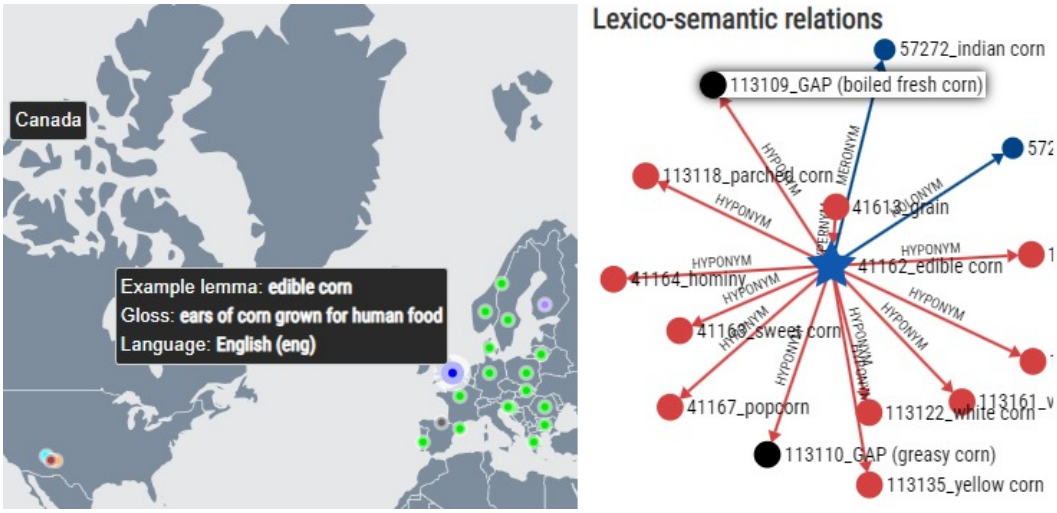


Figure 11.4: Exploring the concept of "*edible corn*" as lexicalized in the English language, in the world (left), and as part of the shared concept hierarchy (right).

describing the concept جَدّ meaning "*parent's father*". On the left-hand side of the screenshot, details are provided on the lexicalization of the concept in Arabic, such as synonymous words, a definition, and a part of speech. The middle part of the screenshot shows an interactive clickable map of all lexicons that either contain the concept or, on the contrary, lack it due to their languages being known not to lexicalize it. The color-coded dots indicate the language family, while the black circled dot represents a lexical gap. This map presents an instant global typological overview of the concept selected; for instance, from Figure 11.2, one can see that most languages in Europe lexicalize the concept جَدّ while several languages in the American United States do not lexicalize it. Finally, the right-hand side shows the concept جَدّ in the context of concept hierarchy, depicted as an interactive graph: the concept, its parent and child concepts, and other lexical-semantic relations (as metonymy and meronymy) are also presented when they exist. Note that the graph only shows a part of the complete hierarchy for usability reasons. Nevertheless, it is navigable and allows the exploration of the whole concept graph in the selected language.

The produced Indonesian, Javanese and Banjarese kinship datasets from the same experiment were also imported into the UKC lexicons. Figure 11.3 shows how UKC explores information about a specific Indonesian word. However, the screenshot provides information about the Indonesian word *saudara*, which means "*sibling*" in English. The left-hand side of the screenshot explains synonymous words (lemmas) and the definition of the typed word. The middle of the screenshot displays the map of a global typologi-

cal overview of the concept. Most languages do not lexicalize this concept, marked by the black-circled dot. Only a few languages lexicalize it, such as Indonesian, Swedish, Ainu, and Malayalam, marked by white-circled dots. The right-hand side shows the lexico-semantic relations of the concept.

Another example is from the food domain. Figure 11.4 shows a screenshot of the concept exploration functionality, which describes the concept *edible corn*, defined as "*ears of corn grown for human food*" in the English language. The left-hand side of the screenshot displays an interactive, clickable map that provides an instant global typological overview of the selected concept, showing all UKC lexicons that either lexicalize the concept or identify it as a lexical gap. On the right-hand side, the concept *edible corn* is displayed within its conceptual hierarchy through an interactive graph. As seen in the previous examples—discussing the Arabic concept جَدّ and the Indonesian concept *saudara*—this graph visualizes not only the concept itself but also its parent and child concepts, along with other lexical-semantic relationships, such as hypernyms, hyponyms, meronyms, and holonyms, where applicable. In addition to showing the lexicalizations of concepts related to *edible corn*, the graph also highlights concepts that represent lexical gaps in English. For instance, *boiled fresh corn*, a hyponym of *edible corn*, is a lexical gap in English but is lexicalized in several Amerindian languages, such as Zuni, Hopi, and Western Keres.

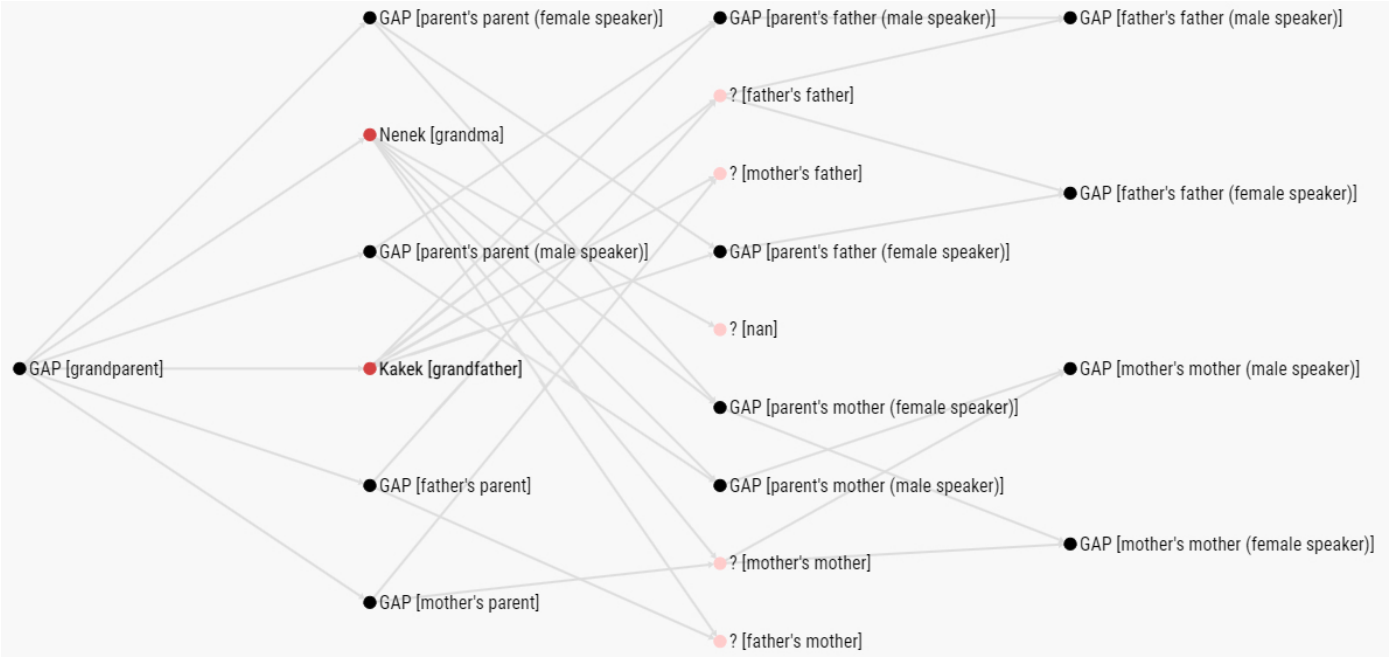


Figure 11.5: Interactive browser tool showing lexical units and gaps for the *grandparent* subdomain in Indonesian

The UKC lexicon is also equipped with several interactive visualization services that can be used to browse lexical units and gaps by domain in all supported languages. Figure 11.5 shows an example of using such services in visualizing the content of the grandparent subdomain in Indonesian.

11.2 Arabic UKC

Arabic UKC[1] is an ongoing project dedicated to developing a UKC instance specifically for Arabic dialects. To our knowledge, it is the first diversity-aware lexical resource designed for Arabic dialects, encompassing words and their meanings from both Standard Arabic and twenty regional dialects. Similar to the main UKC, its structure includes a top layer for language independent concepts and a bottom layer comprising twenty lexicons, each corresponding to one of the Arabic dialects. A screenshot of the Arabic UKC homepage is shown in Figure 11.6. The dropdown list displays the languages included in the Arabic UKC. Additionally, the figure provides details about the selected language, such as the total number of words, word senses, and lexical gaps in its lexicon.

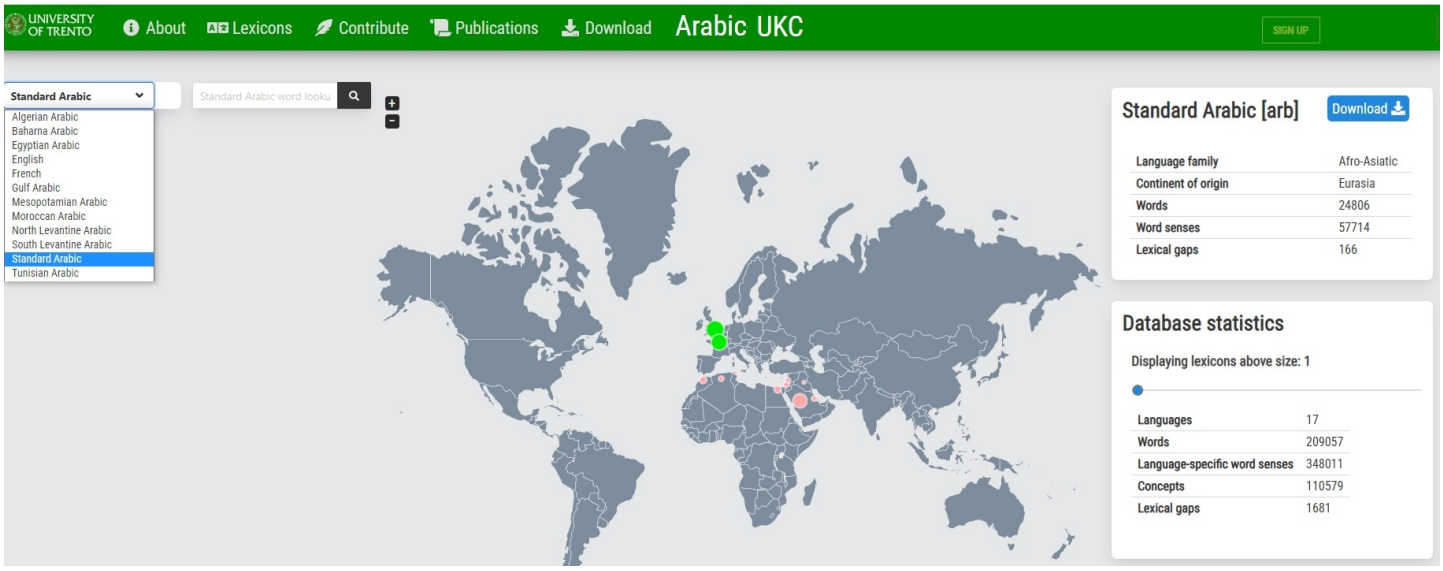


Figure 11.6: Homepage of the Arabic UKC ongoing project

Figure 11.7 shows a screenshot of the concept exploration functionality, illustrating the concept أَبْلَة, which means "*elder sister*". As in the webpage of the main UKC, the left-hand side of the screenshot provides details about the lexicalization of the concept in Egyptian Arabic, including synonymous words, a definition, and its part of speech. The middle section of

[1] http://arabic.ukc.datascientia.eu

the screenshot shows an interactive, clickable map of the Arabic dialect lexicons, highlighting whether a given dialect lexicalizes the concept or leaves it as a lexical gap due to its absence in that dialect. White circled dots indicate lexicalization of the concept, while black circled dots represent lexical gaps. This map offers an instant global typological overview of the selected concept. For example, as shown in Figure 11.7, Gulf and Algerian dialects lexicalize the concept as بنت العود and لالّة, respectively, while Tunisian, Syrian, and Palestinian dialects do not lexicalize it. Finally, the right-hand side displays the concept أَبْلَه within the concept hierarchy as an interactive graph. This graph illustrates the concept along with its parent and child concepts, as well as other lexical-semantic relationships, such as hypernymy and hyponymy, when available.

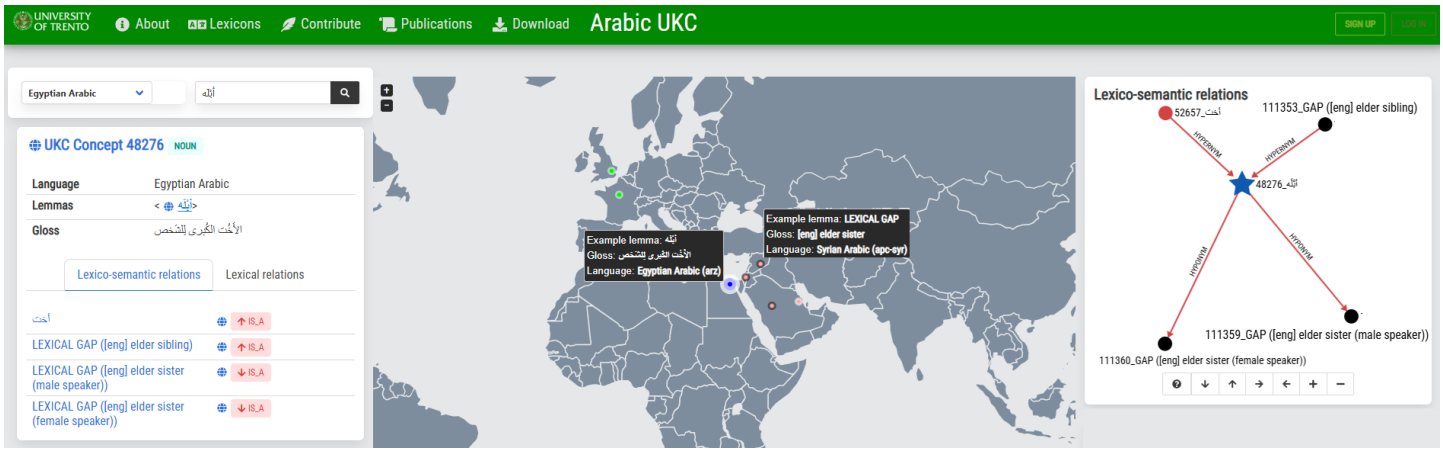


Figure 11.7: Exploring the concept of أَبْلَه as lexicalized in the Egyptian dialect (left), in the Arabic world (middle), and as part of the shared concept hierarchy (right)

11.3 Indonesian UKC

Indonesia is home to over 700 languages, many of which are under-researched, raising concerns about the potential for language extinction [3]. The Indonesian UKC[2] is an ongoing project aimed at developing a dedicated UKC instance for the languages of Indonesia. To the best of our knowledge, it is the first diversity-aware lexical resource designed specifically for these languages. It incorporates words and their meanings from both Bahasa Indonesia and the 700 national languages. Similar to the Arabic UKC, its structure includes a top layer for language-independent concepts and a bottom layer consisting of 700 lexicons, each representing one of Indonesia's languages. A screenshot of the Indonesian UKC homepage is shown in Figure 11.8.

[2] http://indonesia.ukc.datascientia.eu

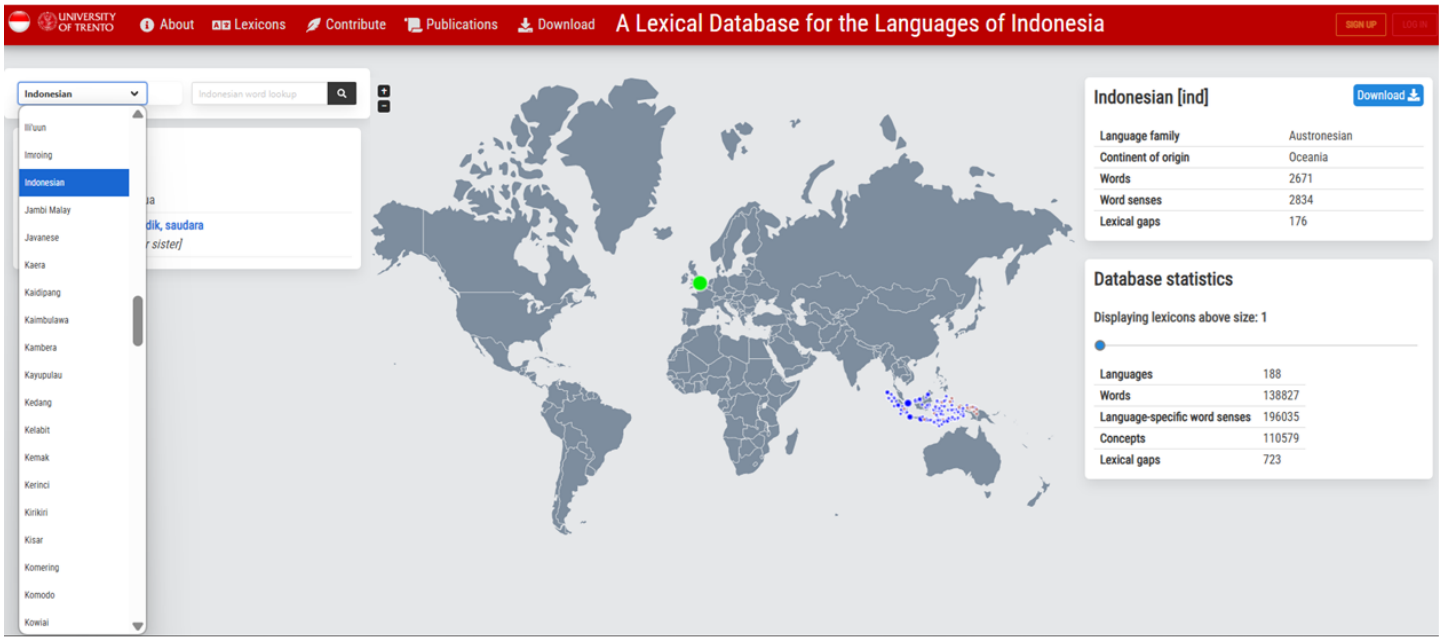


Figure 11.8: Homepage of the Indonesian UKC ongoing project

The dropdown list displays the languages included in the Indonesian UKC. The figure also presents details about the selected language, such as its language family, the total number of words, word senses, and lexical gaps in its lexicon. For instance, when Indonesian is selected from the dropdown list on the left-hand side, the right-hand side displays the following information: the language family is *Austronesian*, there are 2,671 Indonesian words, and 176 lexical gaps are identified in the Indonesian lexicon. Note that the datasets—comprising words and lexical gaps in Indonesian, Banjarese, and Javanese—resulting from the experiments in the food domain and basic-level categories are currently being integrated into the Indonesian UKC.

The lexical units and gaps identified in these Indonesian languages as a result of the kinship experiment have already been incorporated into the Indonesian, Banjarese, and Javanese lexicons of the Indonesian UKC. For instance, a screenshot of the concept exploration functionality is shown in Figure 11.9, illustrating the concept of *kakak*, which means "*elder sibling*". On the left-hand side, the screenshot provides detailed information about the lexicalization of this concept in Indonesian, including its lemmas, gloss, part of speech, and its lexico-semantic relationships with other concepts. The middle section displays an interactive map of Indonesian lexicons, indicating whether a language lexicalizes the concept or considers it a lexical gap due to its absence. Lexicalizations of the concept are represented by white circled dots, while lexical gaps are shown as black circled dots. For example, as shown in Figure 11.9, the concept is lexicalized in Indonesian

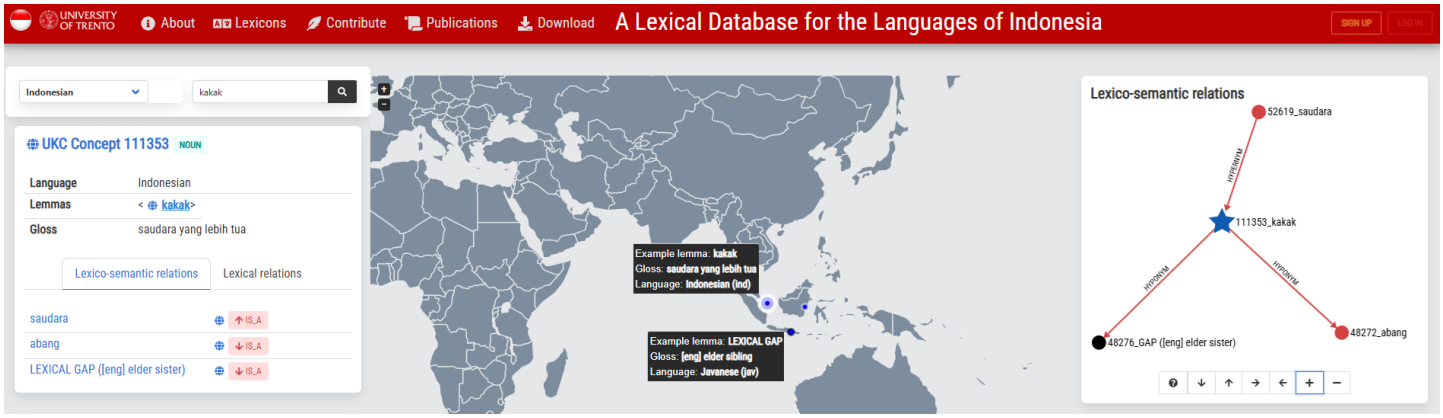


Figure 11.9: Exploring the concept of *kakak* as lexicalized in Indonesian (left), in the Indonesian languages (middle), and as part of the shared concept hierarchy (right).

but not in Javanese. On the right-hand side, an interactive graph visualizes the graphical relationships of the concept *kakak* with its parent and child concepts.

11.4 Arabic WordNet (AWN)

To the best of our knowledge, AWN V3—the latest version of AWN developed through our work—is the first diversity-aware wordnet designed for Modern Standard Arabic (MSA). MSA refers to the standardized form of the language used in academic writing, formal communication, classical poetry, and religious sermons. AWN V3 incorporates 9,576 synsets derived from AWN V1, whose quality has been enhanced in this study by adding missing information and correcting errors in the original synsets. Furthermore, the wordnet structure has been extended with new elements, such as phrasets and lexical gaps, to address issues related to language diversity and untranslatability.

As a result of these efforts, 5,578 synsets (58% of the original 9,576) were updated. Detailed information on these updates is provided in Table 10.5.

Below are examples illustrating synsets before (original AWN V1 synsets) and after updates (AWN V3 synsets). The first example highlights synset enhancement through the deletion and correction of lemmas, as well as the addition of a phraset to a synset. The second example demonstrates the identification of a lexical gap in another synset.

Example 4 *Original vs. improved: synset enhancement*

Original AWN V1 synset: *"جِسْم، شَيْء", corresponding to the English synset: "object, physical object: a tangible and visible entity; an entity that*

can cast a shadow; pens and books are school objects".

Improved AWN V3 Synset: *"*جِسْم فِيزْيَائِي: أجسام ملموسة، يكان ذو ظل. الجاذبية الأرضية تجعل الأجسام المادية تسقط على الأرض*".*

This example illustrates how the original synset from AWN V1 was improved to produce the updated synset included in AWN V3 by performing the following actions:

1. *Deleting the lemma* شَيْء*, which was deemed an incorrect synonym.*
2. *Correcting the lemma* جِسْم *to* جِسْم فِيزْيَائِي*.*
3. *Adding the phraset* جسم مادي *to enhance the clarity and understandability of this synset.*

Example 5 *Original vs. improved: lexical gap identification*

Original AWN V1 Synset: *"*تَعْبِيرِيّاً: بطريقة تعبر عن الشخص بالنسبة لما يفكر أو يشعر به*", corresponding to the English synset: "expressively: with expression, in an expressive manner; she gave the order to the waiter, using her hands very expressively".*

In AWN V3, this synset was identified as a lexical gap, and the phraset بشكل معبر *was added to enhance its representation.*

11.5 Application in machine translation

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

11.5.1 The gap problem in machine translation

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

The sentences in Table 11.1 are real machine translation outputs for the English sentence "*His cousin gave birth to twins*" in four languages[3]: Arabic, Turkish, Persian, and Indonesian.

[3]Obtained from Google Translate on 19 December 2024

Target Language	Target Sentence
Arabic	*أنجب ابن عمه توأما
Turkish	*Kuzeni ikiz doğurmuştu.
Persian	*پسر عمویش دوقلو به دنیا آورد.
Indonesian	sepupunya melahirkan anak kembar.

Table 11.1: Target languages and target sentences for translating "*His cousin gave birth to twins*" using Google Translate.

The only correct output is in Indonesian, where the English term *cousin* has a direct equivalent, "sepupu". In contrast, *cousin* represents a lexical gap in Arabic, Turkish, and Persian, as these languages use more specific terms to describe this relationship. A corpus-statistics-based translation approach often results in significant meaning-level errors. For example, the Arabic term ابن عمه and the Persian term پسرعمو both specifically mean "*father's brother's son*", while the Turkish term ***kuzen*** refers to a "*male cousin*". These distinctions can lead to nonsensical translations, such as "*His father's brother's son gave birth to a twin*" or "*His male cousin gave birth to a twin*". This output, while syntactically correct, is semantically incorrect—implying a male giving birth—due to the lexical gap.

In cases where an equivalent term does not exist, a semantically informed translation method can choose a broader term instead of a narrower one, achieving approximation at the expense of slight information loss—an outcome less critical than inadvertently introducing false information. However, this approach does not work in the present case because "cousin" serves as the root term for the "cousin" subdomain, and using an even broader term, such as *relative*, may not sound appropriate. In such situations, selecting the correct narrower term—whether ابن عمه "*father's brother's son*" or بنت عمه "*father's brother's daughter*" in Arabic, پسرعمو "*father's brother's son*" or دخترعمو "*father's brother's daughter*" in Persian, or ***kuzen*** "*male cousin*" or ***kuzin*** "*female cousin*" in Turkish—can often be inferred from the sentence context by either a human translator or a sophisticated automated method.

The explicit and formal representation of untranslatability, as provided by the UKC resource, can be exploited not only to improve translation systems but also to evaluate their semantic performance on challenging tasks. In the following section, we illustrate this evaluation scenario through three experimental case studies.

11.5.2 Machine translation evaluation method

We describe a *semantic* evaluation measure and method for MT systems, designed to capture meaning-level translation mistakes that conventional metrics (e.g. the BLEU score) are known not to address adequately [163].

The method consists of:

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

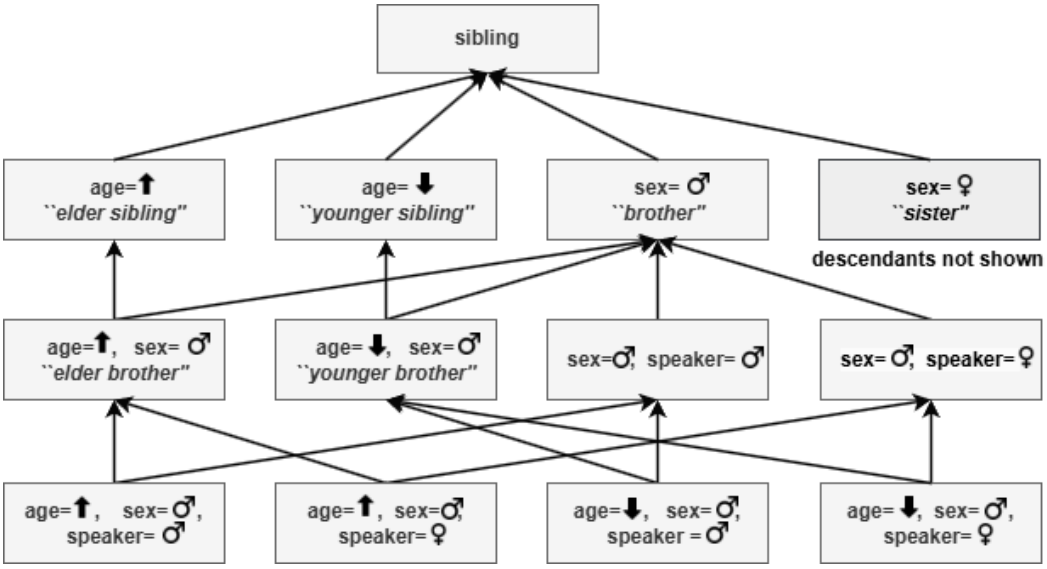


Figure 11.10: Interlingual conceptual layer of the sibling domain

The UKC lexicon, a formal lexical gap resource, is employed in Step 1 for constructing and annotating sentences, and in Step 3 for computing semantic distances based on the *least common subsumer distance* between the original and translated term meanings within the interlingual concept hierarchy.

In Step 1, three case studies were conducted on Arabic, Turkish, and Indonesian, with English serving as the reference language (either as a source or target language). For each of the three languages, a linguistic expert (a university instructor) built a corpus of 60 sentences containing kinship terms. These sentences included 2–13 representative terms from each kinship subdomain, with each term annotated according to its corresponding interlingual concept. Consequently, three corpora were created: 60 English sentences for the Arabic-English study, 60 Turkish sentences for the Turkish-English study, and 60 Indonesian sentences for the Indonesian-English study. Given the well-documented lexical diversity of kinship terms, we consider

these corpora to be an effective evaluation set for evaluating meaning-level challenges in hard-to-translate sentences.

In Step 2, for each case study, we translated the 60 sentences using Google Translate, GPT-4, and DeepL[4]. In the Arabic-English study, the sentences were translated into Arabic, whereas in the Turkish-English and Indonesian-English studies, the sentences were translated into English.

In Step 3, we measured the semantic distance between target words and translation outputs. First, we automatically disambiguated output words (e.g., the Arabic أخ was disambiguated as "brother", formally represented as Br). We then computed the semantic distance between this representation and the gold standard annotation. Figure 11.10 illustrates that the Arabic أخ "*brother*" and أخت "*sister*" are connected through the concept of "sibling". The distance from "brother" to "sister" is two hops, resulting in a semantic distance of 2.

Machine translation experiment results

Table 11.2 shows for each language pair the number of gaps, the average semantic distance over gap-containing sentences, and the average semantic distance over all sentences across the three machine translators.

Language Pair	Gaps	Semantic Distance (Gaps)			Semantic Distance (All)		
		Google	GPT-4	DeepL	Google	GPT-4	DeepL
English-Arabic	14	2.0	1.15	1.31	1.1	0.63	0.72
Turkish-English	12	1.58	1.26	1.71	0.96	0.68	0.97
Indonesian-English	4	1.52	0.36	1.29	1.08	0.20	0.86

Table 11.2: Semantic distance of Google Translate, GPT-4, and DeepL in the three case studies

Machine translation of kinship terms proved to be significantly more robust when using GPT-4 compared to Google Translate or DeepL. For example[5], consider the sentence: "*Jack, my cousin on my mother's side, decided to study abroad*", where "cousin" serves as the representative word. Google Translate and DeepL rendered this sentence in Arabic as جاك، ابن عمي من جهة والدتي، قرر الدراسة في الخارج. However, this translation contains an error: the term ابن عمي meaning "*son of father's brother*" instead of the correct term ابن خالي, which means "*son of mother's brother*". In contrast, ChatGPT-4 accurately translated the sentence as جاك، ابن خالي من جهة أمي، قرر الدراسة في الخارج.

[4]Since Persian is not supported by DeepL, we did not include a case study on Persian in the MT evaluation experiment.

[5]Obtained from Google Translate, GPT-4, and DeepL on 03 January 2025

Another example[6] is "Amcaçocuğu Serkan bugün bize geldi", *Amcaçocuğu* means "*father's brother's child*"—a concept that lacks a direct lexical equivalent in English. Additionally, *Serkan* is a common Turkish boy's name. ChatGPT-4 translated this sentence as: "Serkan, *the son of my uncle*, came to us today", achieving a semantic distance of 1. In contrast, Google Translate rendered it as: "My *cousin* Serkan came to us today", with a semantic distance of 2, while DeepL produced: "His *uncle* Serkan came to us today", resulting in a semantic distance of 5. Accordingly, ChatGPT-4 effectively addressed the lexical gap by selecting the concept most closely aligned with "father's brother's child" within the cousin domain's interlingual conceptual layer.

Furthermore, translations from Indonesian to English demonstrated higher accuracy compared to those from Turkish to English or English to Arabic. This discrepancy can likely be attributed to the smaller number of lexical gaps in the Indonesian-English case study compared to the other languages.

On the whole, it is evident that gap-based distances are consistently higher than the overall distances. This observation highlights that managing lexical gaps remains a significant weakness in current MT techniques—an issue that likely warrants targeted solutions for improvement.

[6]Obtained from Google Translate, GPT-4, and DeepL on 04 January 2025

12 Conclusion

Contents

12.1 Summary

This thesis addresses the critical challenge of linguistic diversity and representation in multilingual lexical-semantic resources (LSRs), which are foundational for natural language processing (NLP) tasks. Traditional LSRs often exhibit English-centric biases and inadequacies in capturing cross-lingual lexical gaps, particularly in low-resource languages. The research focuses on developing diversity-aware LSRs through a hybrid approach consisting of two methodologies: expert-sourcing and crowdsourcing. The work is organized into four key parts: the *first part* of this work, particularly Chapter 2, introduces the concepts of LSRs, highlights their foundational role in NLP, and identifies quality issues in existing LSRs, such as gaps in semantic relationships and incorrect or incomplete entries, especially for low-resource languages. Chapter 3 discusses common limitations of LSRs, including language bias and lexical untranslatability. It emphasizes the need to address linguistic diversity to mitigate lexical gaps and create accurate, comprehensive multilingual resources, while underscoring the significance of cultural and linguistic nuances in NLP.

In Chapter 4, within the *second part*, an expert-driven methodology is proposed to create high-quality, diversity-aware datasets. This approach involves collaborating with linguists and translators to systematically identify lexical gaps and terms across languages. In Chapter 5, case studies on kinship terminology in Arabic dialects and Indonesian languages, as well as experiments on basic-level categories in Arabic, Turkish, Persian, Indonesian, Banjarese, and Javanese, demonstrate the method's effectiveness. The outcomes include a more comprehensive and diversity-aware Arabic Word-

Net (AWN V3) and the enrichment of resources like the UKC with lexical diversity data.

The *third part* focuses on leveraging crowdsourcing to enrich linguistic resources by harnessing collective contributions from diverse populations. Chapter 6 explores the concept, applications, and advantages of crowdsourcing, particularly in creating multilingual and culturally sensitive lexical datasets. Chapter 7 presents a *novel crowdsourcing methodology* designed to systematically collect data on lexical equivalence and gaps across languages. This methodology comprises three key steps: *task generation* (using tools such as databases, language models, or fieldwork to create input data), *task execution* (gathering lexical gaps and words using crowdsourcing platforms like LingoGap), and *task validation* (utilizing inter-rater agreement and expert reviews). Chapter 8 highlights the *LingoGap* platform, describing how it facilitates the crowdsourcing process through user-friendly interfaces and systematic task management. LingoGap serves as a central tool, offering a structured environment for task creation, participant interaction, and quality assurance via attention checks and inter-rater agreement metrics. Finally, Chapter 9 validates the methodology through a series of experiments involving diverse language pairs (e.g., English-Arabic, Indonesian-Banjarese) and semantic fields such as food and kinship. These experiments demonstrate the scalability and adaptability of the crowdsourcing approach. By incorporating measures such as expert reviews and automated agreement checks, the methodology ensures the collection of high-quality data that effectively captures linguistic nuances in both contemporary and underrepresented languages.

Finally, Part Four consists of two chapters. Chapter 10 presents the outcomes of expert-sourcing and crowdsourcing methodologies, which generated diverse datasets for linguistic and cross-linguistic research. These datasets include kinship terms, basic-level categories, and food-related lexical items from various languages, including Arabic, Indonesian, Banjarese, Javanese, Turkish, Persian, and Maba. Additionally, the datasets incorporate lexical gaps and highlight cultural nuances. The chapter emphasizes the accessibility of these datasets through platforms such as DataScientia and provides linguistic insights derived from the collected contributions. Chapter 11 discusses the integration of these datasets into high-quality multilingual lexicons, such as the UKC and the Arabic and Indonesian UKCs, demonstrating their utility in enhancing linguistic resources like the Arabic WordNet and in improving and evaluating machine translation systems.

In conclusion, this thesis advances the state-of-the-art in constructing multilingual LSRs by introducing systematic and scalable methods to capture lexical diversity. The outcomes contribute significantly to improving the quality and inclusiveness of NLP systems, with potential cross-lingual applications in machine translation, word sense disambiguation, sentiment analysis, and other domains.

12.2 Discussion

12.2.1 Benefits and limitations

Benefits

The hybrid approach, which combines expert-sourcing with crowdsourcing, offers several significant benefits. First, it *systematically* identifies and documents linguistic diversity aspects, including *lexical gaps*, *language- and culture-specific concepts*, and *equivalent word meanings*. This approach provides valuable insights into language-specific phenomena. The *direct* involvement of *native speakers* ensures contextual accuracy by capturing nuanced linguistic and cultural meanings in lexical contributions.

Moreover, our crowdsourcing method facilitates the creation of lexical diversity-aware datasets, enabling *bidirectional* lexical exploration between source and target languages across socially significant domains such as food, kinship, and motion verbs. The method demonstrates *scalability* and *adaptability* across both languages and domains. It efficiently handles datasets of varying sizes, partitioning larger domains while remaining flexible enough to accommodate diverse languages, including *low-resource* ones. This adaptability is achieved through the integration of lexical databases, language models, and fieldwork resources. Task efficiency is further enhanced by breaking down complex problems into smaller, manageable micro-tasks, enabling even non-expert contributors (e.g., ordinary speakers) to participate effectively. Live quality control mechanisms, such as ACQs and completion time monitoring, ensure accurate and reliable contributions.

Using Krippendorff's Alpha, two methods were developed to assist researchers in obtaining high-quality contributions through crowdsourcing. The first method is an algorithm for *crowd filtering based on Alpha*, which identifies low-quality crowd workers by calculating IAA during pilot experiments conducted before initiating actual crowdsourcing tasks. The second method involves *validating crowdsourced data using Alpha*, allowing researchers to automatically evaluate contributions based on IAA calculated from multiple participants.

The *LingoGap* platform facilitates these tasks through a user-friendly interface, providing clear instructions and structured workflows to optimize contributor efficiency and ease of use. *Transparency* is another key feature, supported by tools for tracking *task duration* and recording *response timestamps*. These features collectively enhance data accountability and enable more precise analysis, ensuring high-quality contributions.

Additionally, the approach seamlessly *integrates* lexical gaps and units into existing resources, such as the UKC, enriching them with diversity-aware data and enhancing their overall utility. Notably, it has successfully *improved* existing lexical resources, as demonstrated by the enhancements to Arabic WordNet, resulting in the first diversity-aware Arabic lexicon and an

overall improvement in content quality. These features collectively make the hybrid approach as a robust and adaptable solution for exploring linguistic diversity across various domains.

Limitations

The primary limitations of this study arise from the *timing* of task execution in relation to worker availability. Tasks scheduled during *peak periods*—such as exam seasons, holidays like Christmas or Iftar, or times of personal commitments—frequently encounter delays or remain incomplete due to reduced workforce participation. Moreover, our hybrid approach, particularly the crowdsourcing methodology, is highly dependent on *group performance*, rendering delays caused by individual members particularly disruptive to the overall workflow. These disruptions can lead to extended task completion times and decreased efficiency in achieving project milestones.

The *expert-sourcing methodology* hinges on the *availability* and linguistic *expertise* of qualified professionals, which varies considerably across languages—especially in the case of low-resourced languages. For instance, this study required an extensive five-month search to identify a proficient Banjarese linguistic expert for the contribution presented in Section 5.2. Additionally, the *scalability* of the expert-sourcing approach is constrained, particularly when dealing with large datasets or languages lacking robust linguistic resources and suffering from a scarcity of experts or native speakers. This methodology also relies on external linguistic resources, such as dictionaries, typological datasets, and online platforms like Wiktionary. However, these resources are not always comprehensive, accurate, or up-to-date, which may limit the methodology's effectiveness.

One of the key challenges in applying the *crowdsourcing methodology* was recruiting a *substantial number of crowd workers* to participate in our validation experiments. This highlights the need to test the method with an even larger and more diverse crowd to ensure a comprehensive and reliable evaluation. Another significant challenge arises from the difficulty that ordinary language speakers face in determining whether *ambiguous* SL terms—those existing in a *semantic gray area*—should be classified as lexicalized terms or lexical gaps in the TL. Our experiments revealed several instances of such ambiguity. For example, the Arabic word مهلبية "*a type of dessert made from milk, starch, and sugar*" discussed in Section 9.1, the Banjarese word "rabuk" "*a type of ground meat*" in Section 9.2, and the Persian phrase برادر رضاعی "*breastfeeding brother*" in Section 9.7 all illustrate this linguistic uncertainty.

Our study also encountered unique challenges in addressing *newly coined terms* and *rare lexical items*. These terms often lack sufficient documentation in existing resources, complicating their validation and integration into the resulting resource. Unlike lexical gaps, these terms—along with the *new*

language-independent concepts explored in our experiment (e.g., the new kinship concepts identified in Sections 5.1 and 5.2)—pose significant difficulties. Incorporating them into the UKC required the expertise of a lexico-semantic specialist and substantial time, as it sometimes necessitated restructuring the supra-lingual concept hierarchy, as illustrated in Figure 11.1.

Another significant methodological limitation lies in the challenge of capturing *connotative meanings*—those subtle emotional or cultural associations that extend beyond a term's denotative (literal) meaning. While both expert-sourcing and crowdsourcing approaches provide mechanisms for eliciting culturally situated interpretations, they often fall short in systematically capturing affective or metaphorical nuances. Contributors—particularly crowd workers—may interpret tasks (compare source SL words) through personal lenses shaped by individual experience, regional norms, or language ideologies, resulting in variability or inconsistency in semantic judgments [94, 160]. This issue is especially salient for evaluative terms, where meaning is highly context-dependent, and conventional IAA metrics (e.g., Krippendorff's Alpha) may not adequately capture these connotative divergences.

Moreover, *sociolinguistic variation*—differences in language use across regions, social classes, age groups, or genders—presents an additional challenge. While the methodologies employed in this study are effective in identifying lexical gaps and equivalents, they may underrepresent intra-language diversity. For example, a term prevalent in a particular dialect or sociolect may not be widely recognized or may carry different pragmatic implications across subgroups. Since crowdsourcing platforms often attract a digitally literate demographic, there is a risk of overrepresenting certain sociolinguistic profiles while marginalizing others [50]. Consequently, the generated lexical resources may lack comprehensive sociolinguistic depth, underscoring the need for future enhancements that incorporate stratified sampling or sociolinguistically-aware task design.

Finally, we faced challenges in *formalizing* input datasets for *large-scale domains*, such as food. Since the food concept hierarchy in the supra-lingual structure of the UKC—aligned with the food ontology described in Dutta et al. [48]—is extensive, restructuring it to accommodate new food concepts discovered in Sections 9.1 and 9.2 required specialized expertise, significant effort, and time.

12.2.2 Ethics statement

We followed the Guidelines for Research Ethics in the Social Sciences and the Humanities as detailed in Staksrud et al. [141] in this study. The methodologies employed, namely expert-sourcing and crowdsourcing, are interactive approaches that leverage human participation—whether from linguistic experts or crowd workers—to identify and validate lexical gaps and words.

The *ethical considerations* of this research encompass data privacy, integrity, transparency, inclusivity, informed consent, and accountability across both expert-sourcing and crowdsourcing methodologies. Institutional guidelines were strictly adhered to, with approval obtained from the Research Ethics Committee of the *University of Trento*, ensuring compliance with established ethical standards for research involving human participants. Participants, all of whom were at least 20 years old, were fully informed about the study's purpose, nature, and potential outcomes. They provided *explicit consent* before participation and retained the right to withdraw at any time without penalty.

To safeguard privacy, participant identities were *anonymized* throughout the processes of data collection, analysis, and publication. Personal information was neither stored nor shared beyond the study's scope. Data integrity was maintained through rigorous validation procedures, including expert evaluations, Krippendorff's Alpha reliability measures, attention check questions, and completion time metrics, ensuring both the *quality and reliability of contributions*. Furthermore, significant efforts were made to minimize linguistic and cultural biases during data collection, with particular attention given to representing *underrepresented languages* to create balanced and inclusive datasets.

In line with our commitment to *transparency*, the resulting diversity-aware datasets have been made publicly accessible through platforms such as DataScientia. This initiative not only fosters academic collaboration but also ensures *reproducibility* and *openness*, aligning with broader ethical principles of research accountability and inclusivity.

12.3 Future research development

An important direction for future research is extending the hybrid approach introduced in this thesis to encompass a *broader range of languages and dialects*. This is particularly crucial for low-resource languages, which often lack accurate and comprehensive lexical-semantic resources.

Another significant research area involves identifying and documenting lexical gaps in underrepresented cultural domains. This can be achieved by examining *diversity-rich domains* such as body parts [159], colors [127], and visual objects [64], where cultural and linguistic nuances are especially pronounced. Expanding the approach to include these languages and semantic domains would not only enhance linguistic diversity in computational resources but also address the urgent need for inclusivity in multilingual NLP systems.

Developing innovative crowdsourcing models that incorporate *real-time validation* by domain experts (e.g., integrating such services with LingoGap) represents a promising avenue for advancing our crowdsourcing approaches.

Expert feedback during the data collection process can accelerate the production of high-quality lexical resources. Additionally, integrating *gamification* strategies could help expand the participant demographic, such as engaging school students through simple micro-tasks, fostering inclusivity and participation from diverse groups.

Creating a *mobile application* version of the LingoGap platform could further enhance accessibility for the general public, enabling participation in two key areas: (1) developing and enriching existing diversity-aware resources (e.g., the UKC) with lexical units and gaps, and (2) evaluating and improving the quality of lexical content, as demonstrated in our work with the Arabic WordNet.

An intriguing direction for future research involves leveraging diversity-aware lexical resources to *enhance LLMs*. These resources could be employed during the training and testing phases of LLMs to enrich their understanding of linguistic diversity and cultural nuances. Exploring their impact on LLM performance could unveil new possibilities for improving the models' adaptability and effectiveness. Furthermore, designing and evaluating *cross-lingual NLP applications* with enhanced diversity-aware lexical resources could demonstrate the practical utility of this research in real-world scenarios such as machine translation [76], word sense disambiguation [34], and multilingual information retrieval [119]. Finally, future work could also investigate *cross-disciplinary applications* of diversity-aware datasets in fields such as education and cultural studies [116, 137].

Bibliography

[1] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. "On the evaluation and improvement of Arabic WordNet coverage and usability." In: *Language Resources and Evaluation* 47 (2013), pp. 891–917. DOI: 10.1007/s10579-013-9237-0. URL: https://link.springer.com/article/10.1007/s10579-013-9237-0 (cit. on pp. 36, 61).

[2] Arleta Adamska-Sałaciak. "Examining equivalence." In: *International Journal of Lexicography* 23.4 (2010), pp. 387–409 (cit. on pp. 16, 17).

[3] Alham Fikri Aji, Genta Indra Winata, et al. "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 7226–7249 (cit. on pp. 53, 166).

[4] Ken Albala. *Food Cultures of the World Encyclopedia:[4 Volumes]*. Bloomsbury Publishing USA, 2011 (cit. on p. 157).

[5] Meysam Alizadeh, Maël Kubli, et al. "Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks." In: *arXiv preprint arXiv:2307.02179* 101 (2023) (cit. on p. 130).

[6] Musa Alkhalifa and Horacio Rodríguez. "Automatically extending NE coverage of Arabic WordNet using Wikipedia." In: *Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco*. 2009, pp. 23–30 (cit. on p. 61).

[7] Cormac Anderson, Tiago Tresoldi, et al. "A cross-linguistic database of phonetic transcription systems." In: *Yearbook of the Poznan Linguistic Meeting*. Vol. 4. 1. De Gruyter Open. 2018, pp. 21–53 (cit. on p. 27).

[8] Febe Armanios and Bogac Ergene. *Halal food: A history*. Oxford University Press, 2018 (cit. on p. 157).

[9] Aryaman Arora, Adam Farris, Gopalakrishnan R, and Samopriya Basu. "Bhasacitra: Visualising the dialect geography of South Asia." In: *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*. Ed. by Nina Tahmasebi, Adam Jatowt, et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 51–57. DOI: 10.18653/v1/2021.lchange-1.7. URL: https://aclanthology.org/2021.lchange-1.7 (cit. on p. 29).

[10] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics." In: *Computational linguistics* 34.4 (2008), pp. 555–596 (cit. on pp. 87, 199).

[11] Bob Ashley, Joanne Hollows, Steve Jones, and Ben Taylor. *Food and cultural studies*. Routledge, 2004 (cit. on pp. 7, 29, 77).

[12] Rohi Baalbaki. *Al-mawrid Al-qareeb Arabic-English Dictionary*. Dar El Ilm Lilmalayin, lebanon, 2005 (cit. on p. 66).

[13] Badan Pengembangan dan Pembinaan Bahasa. *Kamus Besar Bahasa Indonesia*. Indonesia: Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan, 2017 (cit. on p. 54).

[14] Özge Bakay, Özlem Ergelen, et al. "Turkish WordNet KeNet." In: *Proceedings of the 11th Global Wordnet Conference*. Ed. by Piek Vossen and Christiane Fellbaum. University of South Africa (UNISA): Global Wordnet Association, Jan. 2021, pp. 166–174. URL: https://aclanthology.org/2021.gwc-1.19 (cit. on p. 120).

[15] Balai Bahasa Banjarmasin. *Kamus Bahasa Banjar Dialek Hulu-Indonesia*. Indonesia: Departemen Pendidikan Nasional, Pusat Bahasa, Balai Bahasa Banjarmasin, 2008 (cit. on pp. 54, 58).

[16] Timothy Baldwin, Paul Cook, et al. "How noisy social media text, how diffrnt social media sources?" In: *Proceedings of the sixth international joint conference on natural language processing*. 2013, pp. 356–364 (cit. on p. 24).

[17] Mohamed Barbouch, Suzan Verberne, and Tessa Verhoef. "WN-BERT: Integrating wordnet and BERT for lexical semantics in natural language understanding." In: *Computational Linguistics in the Netherlands Journal* 11 (2021), pp. 105–124 (cit. on pp. 1, 18).

[18] Mohamed Ali Batita and Mounir Zrigui. "The enrichment of Arabic Wordnet antonym relations." In: *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I 18*. Springer. 2018, pp. 342–353 (cit. on p. 61).

[19] Khuyagbaatar Batsuren, Gábor Bella, Fausto Giunchiglia, et al. "Cognet: A large-scale cognate database." In: *ACL 2019 The 57th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*. Association for Computational Linguistics. 2019, pp. 3136–3145 (cit. on p. 3).

[20] Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, Fausto Giunchiglia, et al. "Building the mongolian wordnet." In: *Proceedings of the Tenth Global Wordnet Conference*. Global Wordnet Association. 2019, pp. 238–244 (cit. on p. 3).

[21] Khuyagbaatar Batsuren, Omer Goldman, et al. "UniMorph 4.0: Universal Morphology." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 840–855. URL: https://aclanthology.org/2022.lrec-1.89 (cit. on p. 27).

[22] Gábor Bella, Khuyagbaatar Batsuren, Temuulen Khishigsuren, and Fausto Giunchiglia. "Linguistic Diversity and Bias in Online Dictionaries." In: *Frontiers in African Digital Research*. Ed. by Kroeker Lena. Institute of African Studies, 2022, pp. 173–186 (cit. on pp. 27, 28).

[23] Gábor Bella, Erdenebileg Byambadorj, et al. "Language Diversity: Visible to Humans, Exploitable by Machines." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Valerio Basile, Zornitsa Kozareva, and Sanja Stajner. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 156–165. DOI: 10.18653/v1/2022.acl-demo.15. URL: https://aclanthology.org/2022.acl-demo.15 (cit. on pp. 32, 33).

[24] Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. "Tackling Language Modelling Bias in Support of Linguistic Diversity." In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, pp. 562–572. ISBN: 9798400704505. DOI: 10.1145/3630106.3658925. URL: https://doi.org/10.1145/3630106.3658925 (cit. on pp. 1, 3, 32).

[25] Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. "Towards Bridging the Digital Language Divide." In: *arXiv preprint arXiv:2307.13405* (2023) (cit. on p. 27).

[26] Gábor Bella, Fiona McNeill, et al. "A Major Wordnet for a Minority Language: Scottish Gaelic." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, et al. Marseille, France: European Language Resources Association, May 2020, pp. 2812–2818. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.342 (cit. on p. 37).

[27] Martin Benjamin and Paula Radetzky. "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*. 2014 (cit. on p. 75).

[28] Luisa Bentivogli and Emanuele Pianta. "Looking for lexical gaps." In: *Proceedings of the 9th EURALEX International Congress*. Ed. by Ulrich Heid and Stefan Evert. Institut für Maschinelle Sprachverarbeitung, 2000, pp. 663–669 (cit. on p. 36).

[29] Shahzad Sarwar Bhatti, Xiaofeng Gao, and Guihai Chen. "General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey." In: *Journal of Systems and Software* 167 (2020), p. 110611 (cit. on pp. 70, 71).

[30] Chris Biemann and Valerie Nygaard. "Crowdsourcing wordnet." In: *The 5th International Conference of the Global WordNet Association (GWC-2010)*. 2010 (cit. on p. 75).

[31] Mohamed Mahdi Boudabous, Nouha Chaâben Kammoun, et al. "Arabic WordNet semantic relations enrichment through morpho-lexical patterns." In: *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. IEEE. 2013, pp. 1–6 (cit. on p. 61).

[32] Richard A Brualdi. *Introductory combinatorics*. Pearson Education India, 2004 (cit. on pp. 83, 88).

[33] Samuel Cahyawijaya, Holy Lovenia, et al. "NusaCrowd: Open Source Initiative for Indonesian NLP Resources." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13745–13818. DOI: 10.18653/v1/2023.findings-acl.868. URL: https://aclanthology.org/2023.findings-acl.868 (cit. on p. 111).

[34] Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. "DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4331–4352. DOI: 10.18653/v1/2022.acl-long.298. URL: https://aclanthology.org/2022.acl-long.298 (cit. on pp. 1, 19, 180).

[35] Gerd Carling, Filip Larsson, et al. "Diachronic Atlas of Comparative Linguistics (DiACL)–A database for ancient language typology." In: *PLOS ONE* 13.10 (2018), e0205313. DOI: 10.3389/fnins.2013.12345. URL: http://www.frontiersin.org/Journal/10.3389/fnins.2013.12345/abstract (cit. on p. 27).

[36] John Cunnison Catford. *A Linguistic Theory of Translation*. London: Oxford University Press London, 1965 (cit. on pp. 16, 28).

[37] Nandu Chandran Nair. "A Crowdsourcing Methodology for Improving the Malayalam Wordnet." In: *Available at SSRN 4064783* (2022) (cit. on p. 75).

[38] Nandu Chandran Nair, Rajendran S. Velayuthan, et al. "IndoUKC: A Concept-Centered Indian Multilingual Lexical Resource." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, et al. Marseille, France: European Language Resources Association, June 2022, pp. 2833–2840. URL: https://aclanthology.org/2022.lrec-1.303 (cit. on p. 37).

[39] Evgenia Christoforou, Pinar Barlas, and Jahna Otterbacher. "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. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445317. URL: https://doi.org/10.1145/3411764.3445317 (cit. on p. 98).

[40] Jaka Cibej and Spela Arhar Holdt. "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*. 2019 (cit. on p. 75).

[41] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104. eprint: https://doi.org/10.1177/001316446002000104. URL: https://doi.org/10.1177/001316446002000104 (cit. on p. 199).

[42] Bernard Comrie. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, 1989 (cit. on p. 26).

[43] Paul Cook and Suzanne Stevenson. "Automatically Identifying Changes in the Semantic Orientation of Words." In: *LREC*. 2010 (cit. on p. 73).

[44] Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D. Pawar, eds. *The WordNet in Indian Languages*. 1st. Springer Singapore, 2017 (cit. on p. 37).

[45] Mona Diab. "The feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an English Wordnet." In: *Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo*. 2004 (cit. on p. 59).

[46] Bosheng Ding, Chengwei Qin, et al. "Is GPT-3 a Good Data Annotator?" In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 11173–11195. DOI: 10.18653/v1/2023.acl-long.626. URL: https://aclanthology.org/2023.acl-long.626 (cit. on p. 130).

[47] Matthew S. Dryer and Martin Haspelmath, eds. *WALS Online (v2020.3)*. Zenodo, 2013. DOI: 10.5281/zenodo.7385533. URL: https://doi.org/10.5281/zenodo.7385533 (cit. on pp. 29, 33, 38).

[48] Biswanath Dutta, Usashi Chatterjee, and Devika P. Madalli. "From Application Ontology to Core Ontology." In: *Proceedings of the International Conference on Knowledge Modelling and Knowledge Management (ICKM 2013)*. Bangalore, India, Nov. 2013, pp. 67–80. ISBN: 978-93-5137-765-8 (cit. on p. 178).

[49] David Eberhard, Gary F Simons, and Charles D Fenning. *Ethnologue: Languages of Africa and Europe*. SIL International Publications, 2022 (cit. on p. 53).

[50] Penelope Eckert. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Malden, MA: Blackwell Publishers, 2000 (cit. on p. 178).

[51] Sabri Elkateb, William Black, et al. "Building a WordNet for Arabic." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association, 2006, pp. 29–34 (cit. on pp. 5, 14, 22, 36, 49, 60).

[52] Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. "Towards an integrated crowdsourcing definition." In: *Journal of Information science* 38.2 (2012), pp. 189–200 (cit. on p. 70).

[53] Javier Farreres, Horacio Rodríguez, and Karina Gibert. "Semiautomatic Creation of Taxonomies." In: *COLING-02: SEMANET: Building and Using Semantic Networks*. 2002. URL: https://aclanthology.org/W02-1103 (cit. on p. 60).

[54] Christiane Fellbaum. "A semantic network of English verbs." In: *WordNet: An electronic lexical database* 3 (1998), pp. 153–178 (cit. on p. 13).

[55] Christiane Fellbaum and Piek Vossen. "Challenges for a multilingual wordnet." In: *Language Resources and Evaluation* 46 (2012), pp. 313–326. DOI: 10.1007/s10579-012-9186-z. URL: https://link.springer.com/article/10.1007/s10579-012-9186-z (cit. on pp. 25, 36).

[56] Darja Fier, Ale Tavar, and Toma Erjavec. "sloWCrowd: A crowdsourcing tool for lexicographic tasks." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3471–3475. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1106_Paper.pdf (cit. on p. 75).

[57] Lucy Fortson, Karen Masters, et al. "Galaxy zoo." In: *Advances in machine learning and data mining for astronomy* 2012 (2012), pp. 213–236 (cit. on p. 70).

[58] Abed Alhakim Freihat. "An Organizational Approach to the Polysemy Problem in WordNet." PhD thesis. University of Trento, 2014 (cit. on p. 23).

[59] Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. "Solving Specialization Polysemy in WordNet." In: *Int. J. Comput. Linguistics Appl.* 4 (2013), pp. 29–52 (cit. on p. 23).

[60] Abed Alhakim Freihat, Hadi Khalilia, Gábor Bella, and Fausto Giunchiglia. "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*. Ed. by Hend Al-Khalifa, Kareem Darwish, et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 74–83. URL: https://aclanthology.org/2024.osact-1.9 (cit. on pp. 6, 22).

[61] Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. "Using Crowd Agreement for Wordnet Localization." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, et al. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://aclanthology.org/L18-1074 (cit. on p. 74).

[62] Polona Gantar and Simon Krek. "Slovene lexical database." In: *Natural language processing, multilinguality* (2011), pp. 72–80 (cit. on p. 75).

[63] Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev, and Stéphane Polis. "Universal and macro-areal patterns in the lexicon: A case-study in the perception-cognition domain." In: *Linguistic Typology* 26.2 (2022), pp. 439–487 (cit. on p. 31).

[64] Fausto Giunchiglia and Mayukh Bagchi. *Classifying concepts via visual properties*. 2021. arXiv: 2105.09422 [cs.AI]. URL: https://arxiv.org/abs/2105.09422 (cit. on p. 179).

[65] Fausto Giunchiglia, Khuyagbaatar Batsuren, Gabor Bella, et al. "Understanding and Exploiting Language Diversity." In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2017, pp. 4009–4017 (cit. on pp. 1, 3, 4, 17, 24, 27, 33, 38).

[66] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. "One World–Seven Thousand Languages." In: *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018*. Ed. by Alexander Gelbukh. Springer, 2018, pp. 18–24 (cit. on pp. 33, 37).

[67] Fausto Giunchiglia, Gábor Bella, et al. "Representing interlingual meaning in lexical databases." In: *Artificial Intelligence Review* 56.10 (2023), pp. 11053–11069 (cit. on pp. 1, 17).

[68] Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, Khuyagbaatar Batsuren, et al. "Crowdsourcing a large scale multilingual lexico-semantic resource." In: *AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)*. 2015 (cit. on p. 75).

[69] Pius ten Hacken. "Bilingual dictionaries and theories of word meaning." In: *Proceedings of the XVII EURALEX International Congress, Lexicographic Centre, Ivane Javakhishvili Tbilisi State University Tbilisi*. 2016, pp. 61–76 (cit. on p. 17).

[70] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. "Creating language resources for under-resourced languages: methodologies, and experiments with Arabic." In: *Language Resources and Evaluation* 49 (2015), pp. 549–580 (cit. on p. 75).

[71] Maram Hasanain, Fatema Ahmad, and Firoj Alam. *Large Language Models for Propaganda Span Annotation*. 2024. arXiv: 2311.09812 [cs.CL]. URL: https://arxiv.org/abs/2311.09812 (cit. on p. 130).

[72] Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. "Diversity and language technology: how language modeling bias causes epistemic injustice." In: *Ethics and Information Technology* 26.8 (Jan. 2024). DOI: 10.1007/s10676-023-09742-6. URL: https://doi.org/10.1007/s10676-023-09742-6 (cit. on pp. 2, 3, 27, 31, 37).

[73] Jeff Howe. "Crowdsourcing: A definition." In: (2006) (cit. on p. 69).

[74] Oana Inel, Khalid Khamkham, et al. "Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data." In: *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*. Springer. 2014, pp. 486–504 (cit. on p. 70).

[75] Madhuri P Karnik and DV Kodavade. "Homonym Detection Using WordNet and Modified Lesk Approach for English Language." In: *Journal of Electrical Systems* 20.1s (2024), pp. 329–337 (cit. on p. 18).

[76] Akihiro Katsuta and Kazuhide Yamamoto. "Lexical simplification by unsupervised machine translation." In: *International Journal of Asian Language Processing* 30.02 (2020), p. 2050008 (cit. on pp. 1, 19, 180).

[77] Paul Kay and Richard S. Cook. "World Color Survey." In: *Encyclopedia of Color Science and Technology*. Ed. by Ming Ronnier Luo. New York: Springer, 2016, pp. 1265–1271 (cit. on p. 29).

[78] Pei Ke, Haozhe Ji, et al. "SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6975–6988. DOI: 10.18653/v1/2020.emnlp-main.567. URL: https://aclanthology.org/2020.emnlp-main.567 (cit. on p. 19).

[79] Charles Kemp and Terry Regier. "Kinship Categories Across Languages Reflect General Communicative Principles." In: *Science* 336.6084 (2012), pp. 1049–1054. DOI: 10.1126/science.1218811. URL: https://www.science.org/doi/abs/10.1126/science.1218811 (cit. on pp. 29, 55).

[80] Hadi Khalilia, Gábor Bella, et al. "Lexical diversity in kinship across languages and dialects." In: *Frontiers in Psychology* 14 (2023). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2023.1229697. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1229697 (cit. on pp. 2, 3, 5, 24, 31, 81).

[81] Hadi Khalilia, Abed Alhakim Freihat, and Fausto Giunchiglia. "The Dimensions of Lexical Semantic Resource Quality." In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, Dec. 2021, pp. 15–21. URL: https://aclanthology.org/2021.nsurl-1.3 (cit. on pp. 1, 8, 20, 21).

[82] Hadi Khalilia, Abed Alhakim Freihat, and Fausto Giunchiglia. "The quality of lexical semantic resources: A survey." In: *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. 2021, pp. 117–129 (cit. on pp. 1, 8, 19, 22).

[83] Hadi Khalilia, Jahna Otterbacher, et al. *Crowdsourcing Lexical Diversity*. 2024. arXiv: 2410.23133 [cs.CL]. URL: https://arxiv.org/abs/2410.23133 (cit. on pp. 1–3, 6, 32).

[84] Temuulen Khishigsuren, Gábor Bella, et al. "Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, et al. Marseille, France: European Language Resources Association, June 2022, pp. 2798–2807. URL: https://aclanthology.org/2022.lrec-1.299 (cit. on pp. 2, 3, 5, 7, 31–33, 42, 55, 77).

[85] Adam Kilgarriff. "I don't believe in word senses." In: *Computers and the Humanities* 31 (1997), pp. 91–113 (cit. on p. 24).

[86] Kathryn R Kirby, Russell D Gray, et al. "D-PLACE: A global database of cultural, linguistic and environmental diversity." In: *PLOS ONE* 11.7 (2016), e0158391. DOI: 10.1371/journal.pone.0158391. URL: https://doi.org/10.1371/journal.pone.0158391 (cit. on p. 29).

[87] Anetta Kopecka and Bhuvana Narasimhan. *Events of putting and taking: A crosslinguistic perspective*. John Benjamins Publishing, 2012 (cit. on pp. 29, 77).

[88] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. DOI: 10.18653/v1/2020.coling-main.66. URL: https://aclanthology.org/2020.coling-main.66 (cit. on p. 110).

[89] Klaus Krippendorff. *Computing Krippendorff's alpha-reliability*. 2011 (cit. on pp. 87, 199).

[90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 70).

[91] Ewald Lang. "Spatial dimension terms." In: *Language Typology and Language Universals*. Vol. 2. Berlin, Boston: De Gruyter Mouton, 2001, pp. 1251–1275. ISBN: 9783110194265. DOI: doi:10.1515/9783110194265-028. URL: https://doi.org/10.1515/9783110194265-028 (cit. on p. 77).

[92] Bettina Lanser, Christina Unger, and Philipp Cimiano. "Crowdsourcing Ontology Lexicons." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, et al. Portoro, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3477–3484. URL: https://aclanthology.org/L16-1554 (cit. on p. 75).

[93] Adrienne Lehrer. "Notes on Lexical Gaps." In: *Journal of Linguistics* 6.2 (1970), pp. 257–261. URL: https://www.jstor.org/stable/4175082 (cit. on p. 29).

[94] Stephen C Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000 (cit. on p. 178).

[95] Stephen C Levinson and David P Wilkins. *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press, 2006 (cit. on p. 29).

[96] Zefeng Lin, Weidong Chen, Yan Song, and Yongdong Zhang. "Prompting Few-shot Multi-hop Question Generation via Comprehending Type-aware Semantics." In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, pp. 3730–3740 (cit. on pp. 85, 94).

[97] Chen Ling, Xujiang Zhao, et al. "Domain specialization as the key to make large language models disruptive: A comprehensive survey." In: *arXiv preprint arXiv:2305.18703* (2023) (cit. on p. 24).

[98] Johann-Mattis List, Michael Cysouw, and Robert Forkel. "Concepticon: A Resource for the Linking of Concept Lists." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association, 2016, pp. 2393–2400 (cit. on pp. 31, 35, 38).

[99] Qiang Liu, Alexander T Ihler, and Mark Steyvers. "Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?" In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/cc1aa436277138f61cda703991069eaf-Paper.pdf (cit. on pp. 82, 87).

[100] Daniel Loureiro and Alípio Jorge. "Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5682–5691. DOI: 10.18653/v1/P19-1569. URL: https://aclanthology.org/P19-1569 (cit. on pp. 1, 19).

[101] Bernardo Magnini and Gabriela Cavaglià. "Integrating Subject Field Codes into WordNet." In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Ed. by M. Gavrilidou, G. Carayannis, et al. Athens, Greece: European Language Resources Association (ELRA), May 2000. URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/219.pdf (cit. on pp. 37, 60).

[102] Asifa Majid, Melissa Bowerman, Miriam van Staden, and James S Boster. "The semantic categories of cutting and breaking events: A crosslinguistic perspective." In: *Cognitive Linguistics* 18.2 (2007), pp. 133–152. DOI: 10.1515/COG.2007.005. URL: https://www.degruyter.com/document/doi/10.1515/COG.2007.005/html (cit. on pp. 29, 77).

[103] Sanjana Manerkar, Kavita Asnani, et al. "Konkani WordNet: Corpus-Based Enhancement using Crowdsourcing." In: *Transactions on Asian and Low-Resource Language information Processing* 21.4 (2022), pp. 1–18 (cit. on p. 75).

[104] Arya D McCarthy, Winston Wu, et al. "Modeling color terminology across thousands of languages." In: *arXiv preprint arXiv:1910.01531* (2019) (cit. on pp. 24, 27, 29, 42, 77).

[105] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781 (cit. on p. 111).

[106] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41 (cit. on pp. 1, 13, 14, 33, 36).

[107] Saif M Mohammad and Peter D Turney. "Crowdsourcing a word–emotion association lexicon." In: *Computational intelligence* 29.3 (2013), pp. 436–465 (cit. on p. 93).

[108] Mehdi Montazery and Heshaam Faili. "Automatic Persian WordNet Construction." In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). 2010, pp. 2916–2919 (cit. on p. 115).

[109] Briana B Morrison and Betsy DiSalvo. "Khan Academy gamifies computer science." In: *Proceedings of the 45th ACM technical symposium on Computer science education*. 2014, pp. 39–44 (cit. on p. 70).

[110] Christopher Moseley. *Atlas of the World's Languages in Danger*. Unesco, 2010 (cit. on p. 73).

[111] George Peter Murdock. "Kin Term Patterns and Their Distribution." In: *Ethnology* 9.2 (1970), pp. 165–208. DOI: 10.2307/3772782. URL: https://www.jstor.org/stable/3772782 (cit. on pp. 5, 26, 29, 42, 54, 55).

[112] Zainal Muttaqin. "Fiqh Lughah dalam Literatur Arab Klasik." In: *Afaq 'Arabiyah: Jurnal Kebahasaaraban dan Pendidikan Bahasa Arab* 4.2 (2009), pp. 107–122 (cit. on p. 50).

[113] Roberto Navigli. "Word sense disambiguation: A survey." In: *ACM computing surveys (CSUR)* 41.2 (2009), pp. 1–69 (cit. on p. 24).

[114] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: Building a very large multilingual semantic network." In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. 2010, pp. 216–225 (cit. on p. 17).

[115] Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. "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, 2011, pp. 255–264 (cit. on p. 36).

[116] Masatsugu Ono, Toshioki Soga, Masato Kikuchi, and Tetsu Tanabe. "Consideration of Language Learning Service with Visualized Vocabulary Map Derived from WordNet." In: *2023 8th International Conference on Business and Industrial Research (ICBIR)*. IEEE. 2023, pp. 1194–1198 (cit. on pp. 18, 180).

[117] Noam Ordan and Shuly Wintner. "Hebrew WordNet: a Test Case of Aligning Lexical Databases across Languages." In: *International Journal of Translation* 19.1 (2007), pp. 39–58 (cit. on p. 36).

[118] Sam Passmore, Wolfgang Barth, et al. "Kinbank: A global database of kinship terminology." In: *PLOS ONE* 18.5 (2023), e0283218. DOI: 10.1371/journal.pone.0283218. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283218 (cit. on p. 31).

[119] Carol Peters, Martin Braschler, and Paul Clough. *Multilingual information retrieval: From research to practice*. Springer, 2012 (cit. on p. 180).

[120] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. "Developing an aligned multilingual database." In: *Proceedings of the 1st International WordNet Conference*. Global Wordnet Association, 2002, pp. 293–302 (cit. on p. 36).

[121] Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. *A Wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław, 2009 (cit. on p. 14).

[122] Maciej Piasecki, Stan Szpakowicz, Christiane Fellbaum, and Bolette Sandford Pedersen. "Introduction to the special issue: On wordnets and relations." In: *Language Resources and Evaluation* 47 (2013), pp. 757–767 (cit. on p. 25).

[123] VA Plungyan. "Modern linguistic typology." In: *Herald of the Russian Academy of Sciences* 81.2 (2011), pp. 101–113. DOI: 10.1134/S1019331611020158. URL: https://link.springer.com/article/10.1134/S1019331611020158 (cit. on p. 29).

[124] David Martin Ward Powers. "The Problem with Kappa." In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Walter Daelemans. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 345–355. URL: https://aclanthology.org/E12-1035 (cit. on pp. 87, 199).

[125] Michael Rawson, Samuel Dooley, Mithun Bharadwaj, and Rishabh Choudhary. "Topological data analysis for word sense disambiguation." In: *arXiv preprint arXiv:2203.00565* (2022) (cit. on p. 24).

[126] Tatiana Reznikova, Ekaterina Rakhilina, and Anastasia Bonch-Osmolovskaya. "Towards a typology of pain predicates." In: *Linguistics* 50.3 (2012), pp. 421–465 (cit. on pp. 27, 77).

[127] Debi Roberson, Jules Davidoff, Ian RL Davies, and Laura R Shapiro. "Color categories: Evidence for the cultural relativity hypothesis." In: *Cognitive Psychology* 50.4 (2005), pp. 378–411. DOI: 10.1016/j.cogpsych.2004.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0010028504000763 (cit. on pp. 29, 179).

[128] Jonathan Robinson, Cheskie Rosenzweig, Aaron J Moss, and Leib Litman. "Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool." In: *PloS one* 14.12 (2019), e0226394 (cit. on pp. 82, 87).

[129] Horacio Rodríguez, David Farwell, et al. "Arabic WordNet: Semi-automatic Extensions using Bayesian Inference." In: *LREC*. 2008 (cit. on p. 61).

[130] Eleanor Rosch. "Principles of categorization." In: *Cognition and categorization/Erlbaum* (1978) (cit. on p. 57).

[131] Eleanor Rosch, Carolyn B Mervis, et al. "Basic objects in natural categories." In: *Cognitive psychology* 8.3 (1976), pp. 382–439 (cit. on pp. 7, 57).

[132] Christoph Rzymski, Tiago Tresoldi, et al. "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies." In: *Scientific Data* 7.1 (2020), pp. 1–13. DOI: 10.1038/s41597-019-0341-x. URL: https://www.nature.com/articles/s41597-019-0341-x (cit. on p. 27).

[133] Marta Sabou, Kalina Bontcheva, and Arno Scharl. "Crowdsourcing research opportunities: lessons from natural language processing." In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*. 2012, pp. 1–8 (cit. on p. 73).

[134] Elizabeth Salesky, Eleanor Chodroff, et al. "A Corpus for Large-Scale Phonetic Typology." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4526–4546. DOI: 10.18653/v1/2020.acl-main.415. URL: https://aclanthology.org/2020.acl-main.415 (cit. on p. 29).

[135] Hasan M Sayeed, Trupti Mohanty, and Taylor D Sparks. "Annotating materials science text: A semi-automated approach for crafting outputs with gemini pro." In: *Integrating Materials and Manufacturing Innovation* 13.2 (2024), pp. 445–452 (cit. on p. 130).

[136] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005 (cit. on p. 61).

[137] Lasmi Siahaan, Vanny Wiranata, Kamarudin Zai, and Jamaluddin Nasution. "Keterampilan Membaca Pada Pengajaran BIPA Menggunakan Media Digitalisasi." In: *Journal of Science and Social Research* 6.1 (2023), pp. 160–165 (cit. on pp. 18, 180).

[138] Catherine Arnott Smith and Paul J Wicks. "PatientsLikeMe: Consumer health vocabulary as a folksonomy." In: *AMIA annual symposium proceedings*. Vol. 2008. American Medical Informatics Association. 2008, p. 682 (cit. on p. 70).

[139] James Sneddon. *The Indonesian Language*. Sydney: University of New South Wales Press Ltd, 2003 (cit. on pp. 154, 158).

[140] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Ed. by Mirella Lapata and Hwee Tou Ng. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. URL: https://aclanthology.org/D08-1027 (cit. on p. 74).

[141] Elisabeth Staksrud, Ivar Kolstad, et al. *Guidelines for Research Ethics in the Social Sciences and the Humanities*. May 2022. ISBN: 978-82-7682-102-4 (cit. on p. 178).

[142] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: A large ontology from Wikipedia and Wordnet." In: *Journal of Web Semantics* 6.3 (2008), pp. 203–217 (cit. on p. 61).

[143] Jie Tao and Lina Zhou. "A weakly supervised WordNet-Guided deep learning approach to extracting aspect terms from online reviews." In: *ACM Transactions on Management Information Systems (TMIS)* 11.3 (2020), pp. 1–22 (cit. on p. 19).

[144] NLLB Team, Marta R. Costa-jussà, et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672 (cit. on p. 111).

[145] Kaitlyn Teske. "Duolingo." In: *Calico Journal* 34.3 (2017), pp. 393–401 (cit. on p. 70).

[146] Tim Penyusun Kamus Pusat Bahasa. *Kamus Bahasa Indonesia*. Pusat Bahasa Departemen Pendidikan Nasional, 2008 (cit. on p. 110).

[147] Petter Törnberg. *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. 2023. arXiv: 2304.06588 [cs.CL]. URL: https://arxiv.org/abs/2304.06588 (cit. on p. 130).

[148] Dan Tufis, Dan Cristea, and Sofia Stamou. "BalkaNet: Aims, methods, results and perspectives. a general overview." In: *Romanian Journal of Information science and technology* 7.1-2 (2004), pp. 9–43 (cit. on p. 60).

[149] Sutrisno Sastro Utomo. *Kamus Indonesia-Jawa*. Jakarta: PT Gramedia Pustaka Utama, 2015 (cit. on p. 54).

[150] Sandra A Vannoy and Prashant Palvia. "The social influence model of technology adoption." In: *Communications of the ACM* 53.6 (2010), pp. 149–153 (cit. on p. 74).

[151] Åke Viberg. "The verbs of perception: A typological study." In: *Linguistics* 21.1 (1983), pp. 123–162. DOI: 10.1515/ling.1983.21.1.123. URL: https://www.degruyter.com/document/doi/10.1515/ling.1983.21.1.123/html (cit. on pp. 31, 77).

[152] Sonakshi Vij, Devendra Tayal, and Amita Jain. "A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs." In: *Wireless Personal Communications* 111 (2020), pp. 1271–1282 (cit. on p. 18).

[153] Piek Vossen. "A multilingual database with lexical semantic networks." In: *Dordrecht: Kluwer Academic Publishers. doi* 10 (1998), pp. 978–94 (cit. on p. 60).

[154] PJTM Vossen. "Eurowordnet." In: (1999) (cit. on pp. 17, 60).

[155] Bernhard Wälchli and Michael Cysouw. "Lexical typology through similarity semantics: Toward a semantic map of motion verbs." In: *Linguistics* 50.3 (2012), pp. 671–710 (cit. on pp. 27, 77).

[156] Matthijs J Warrens. "Cohen's kappa is a weighted average." In: *Statistical Methodology* 8.6 (2011), pp. 473–484 (cit. on pp. 87, 199).

[157] Johannes Welbl, Nelson F Liu, and Matt Gardner. "Crowdsourcing multiple choice science questions." In: *arXiv preprint arXiv:1707.06209* (2017) (cit. on pp. 85, 94).

[158] Enam Al-Wer. "Arabic Languages, Variation in." In: *Concise Encyclopedia of Languages of the World*. Ed. by Keith Brown and Sarah Ogilvie. Oxford: Elsevier Ltd., 2008, pp. 53–56 (cit. on p. 154).

[159] Anna Wierzbicka. "Bodies and their parts: An NSM approach to semantic typology." In: *Language Sciences* 29.1 (2007), pp. 14–65 (cit. on pp. 7, 29, 77, 179).

[160] Anna Wierzbicka. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK, 1996 (cit. on p. 178).

[161] Indeewari Wijesiri, Malaka Gallage, et al. "Building a WordNet for Sinhala." In: *Proceedings of the Seventh Global Wordnet Conference*. Ed. by Heili Orav, Christiane Fellbaum, and Piek Vossen. Tartu, Estonia: University of Tartu Press, Jan. 2014, pp. 100–108. URL: https://aclanthology.org/W14-0114 (cit. on p. 74).

[162] Jason D Williams, Antoine Raux, and Matthew Henderson. "The dialog state tracking challenge series: A review." In: *Dialogue & Discourse* 7.3 (2016), pp. 4–33 (cit. on p. 74).

[163] Y. Wu, M. Schuster, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." In: *arXiv preprint arXiv:1609.08144* (2016) (cit. on p. 170).

[164] Yongjing Yin, Jiali Zeng, et al. "LexMatcher: Dictionary-centric Data Collection for LLM-based Machine Translation." In: *arXiv preprint arXiv:2406.01441* (2024) (cit. on p. 19).

[165] Omar F Zaidan and Chris Callison-Burch. "Arabic Dialect Identification." In: *Computational Linguistics* 40.1 (2014), pp. 171–202. DOI: 10.1162/COLI_a_00169. URL: https://direct.mit.edu/coli/article/40/1/171/1458/Arabic-Dialect-Identification (cit. on p. 153).

Appendices

A

Calculating Krippendorff's Alpha

A.1 Introduction

Inter-Annotator Agreement (IAA) mechanisms, such as Cohen's Kappa [156] and Krippendorff's Alpha [41, 89], are crucial for validating crowdsourced contributions, especially in tasks involving multiple annotators or contributors. These mechanisms evaluate the level of agreement among annotators regarding their judgments or annotations, providing a measure of reliability. They are widely utilized in computational linguistics, including for assessing the reliability of coders in corpus creation tasks [10]. Krippendorff's Alpha, in particular, is a reliability coefficient that measures agreement among multiple annotators across various data types, including nominal, ordinal, interval, and ratio scales, whereas Cohen's Kappa is limited to nominal data and is applicable to two annotators only [124].

In this thesis, we recruited three to five annotators for each crowdsourcing experiment to perform the assigned tasks. Consequently, we employed the Alpha (α) metric, as presented in Equation (A.1), during the validation step to calculate the reliability among the annotators.

$$\alpha = \frac{P_a - P_e}{1 - P_e} \tag{A.1}$$

Where:

- P_a: The observed weighted percent agreement.
- P_e: The chance (expected) weighted percent agreement.

A.2 Example: step-by-step calculation of alpha

In this example, we demonstrate the application of the equation and calculate Alpha (α) for a data sample derived from the outputs of a pilot experiment conducted on kinship terms between English (as the source language, SL) and Arabic (as the target language, TL). The sample includes

responses from four annotators regarding five English kinship terms: *cousin*, *grandchild*, *great-grandchild*, *grandfather*, and *granddaughter*. In this experiment, these terms are compared with their Arabic counterparts using the *LingoGap* tool, and the results are presented in a spreadsheet. The columns of the spreadsheet are shown in Table A.1.

ID	Worker-id	SL	TL	TL definition	Transaction duration (Sec)	Transaction date
1	76542-00917-34569-24311-88456	cousin	GAP	no target lemma	90	13-11-2023T18:33:42
2	76542-00917-34569-24311-88456	grandchild	حَفِيد	وَلد الْوَلِد أو وَلَدَ الْبِنت	140	13-11-2023T18:36:02
3	76542-00917-34569-24311-88456	great grandchild	GAP	no target lemma	115	13-11-2023T18:37:57
4	76542-00917-34569-24311-88456	grandfather, gramps, granddad, grandad, granddaddy, grandpa	جَدّ	أبو الأب أو أبو الأم	80	13-11-2023T18:39:17
5	76542-00917-34569-24311-88456	granddaughter	حَفِيدَة	بنت الابنة	160	13-11-2023T18:41:57
6	00891-65431-88832-55990-22331	cousin	GAP	no target lemma	125	13-11-2023T18:44:02
7	00891-65431-88832-55990-22331	grandchild	حَفِيد	وَلد والْوَلِد أو وَلد البِنْت	137	13-11-2023T18:46:19
8	00891-65431-88832-55990-22331	great grandchild	GAP	no target lemma	95	13-11-2023T18:47:54
9	00891-65431-88832-55990-22331	grandfather, gramps, granddad, grandad, granddaddy, grandpa	سيد	أبو الأب أو أبو الأم	90	13-11-2023T18:49:29
10	00891-65431-88832-55990-22331	granddaughter	الحفيدة	بنت الابنة	87	13-11-2023T18:50:56
11	71234-98761-43210-78910-11123	cousin	GAP	no target lemma	115	13-11-2023T18:52:51
12	71234-98761-43210-78910-11123	grandchild	Do not know	not answered	132	13-11-2023T18:55:03
13	71234-98761-43210-78910-11123	great grandchild	GAP	no target lemma	150	13-11-2023T18:57:33
14	71234-98761-43210-78910-11123	grandfather, gramps, granddad, grandad, granddaddy, grandpa	الجد	أبو الأب أو أبو الأم	101	13-11-2023T18:59:14
15	71234-98761-43210-78910-11123	granddaughter	الحفيدة	بنت الابنة	118	13-11-2023T19:01:12
16	14132-20213-39387-76540-79100	cousin	GAP	no target lemma	97	13-11-2023T19:02:49
17	14132-20213-39387-76540-79100	grandchild	GAP	no target lemma	122	13-11-2023T19:04:51
18	14132-20213-39387-76540-79100	great grandchild	GAP	no target lemma	133	13-11-2023T19:07:04
19	14132-20213-39387-76540-79100	grandfather, gramps, granddad, grandad, granddaddy, grandpa	الجد	أبو الأب أو أبو الأم	147	13-11-2023T19:09:31
20	14132-20213-39387-76540-79100	granddaughter	GAP	no target lemma	152	13-11-2023T19:12:03

Table A.1: Output sample (raw data)

Computing Alpha (α) involves several steps, as described below:

Step 1: data cleaning and preprocessing

First, we start with the raw data collected from the annotators. Table A.1 presents details about the responses provided by each of the four suspect accounts (annotators) for each of the five English words. Each row in the table includes the worker ID, the English word displayed in the SL column, the worker's response (which could be the Arabic equivalent word along with its definition, a lexical gap, or "Don't know"), the time taken (in seconds) from opening to submitting the response (completion time), and the date of the response.

In this step, we exclude rows with excessively long or short completion times compared to the average completion time (118.20 seconds). Additionally, we remove all rows associated with workers who fail the Attention Check Questions (ACQs). For the purposes of this sample, we assume that all completion times fall within the normal range and that all four annotators have passed the ACQs.

Next, we categorize the responses into three groups: "GAP", "Non-GAP", and "Do not know". As a result, we produce a table containing cleaned and categorized data, as shown in Table A.2.

	76542-00917-34569-24311-88457	**00891-65431-88832-55990-22333**	**71234-98761-43210-78910-11125**	**14132-20213-39387-76540-79101**
Cousin	GAP	GAP	GAP	GAP
Grandchild	Non-GAP	Non-GAP	Do not know	GAP
Great Grandchild	GAP	GAP	GAP	GAP
Grandfather, Gramps, Granddad, Granddaddy, Grandpa	Non-GAP	Non-GAP	Non-GAP	Non-GAP

Table A.2: Cleaned and categorized data

Step 2: building the agreement table

Next, we construct an agreement table that shows the frequency with which each English word (listed in rows) received each rating within the same category (presented in columns), as illustrated in Table A.3. The values r_{ik} in this table represent the number of times annotators assigned the rating category k to the word i. For example, consider word 3, which received the "GAP" rating four times, denoted as $r_{3,1} = 4$.

In this step, we also calculate $\overline{r}$, the average number of annotators who rated each English word. Intuitively, this is simply the total number of

Word(i)	**Rating Category (K)**			
	GAP	**Non-GAP**	**Do not know**	**Total (r_i)**
cousin	4	0	0	4
grandchild	1	2	1	4
great grandchild	4	0	0	4
grandfather, gramps, granddad, grandad, granddaddy, grandpa	0	4	0	4
granddaughter	1	3	0	4
Average $\bar{r}$	2	1.8	0.2	4

Table A.3: Agreement table

ratings divided by the total number of English words, as shown in Equation (A.2). Mathematically, it is expressed as:

$$\overline{r} = \frac{1}{n} \sum_{i=1}^{n} r_i \tag{A.2}$$

Where n is the total number of English words.

$$\overline{r} = 4$$

We will use this number ($\overline{r}$) in step 4, calculating p_a.

Step 3: picking the weight function

In many cases, not all categories hold equal importance. The weight function allows for assigning different weights to various categories based on their significance. This approach is particularly useful when working with ordinal or interval data, where the magnitude of disagreement between categories can vary. However, in our project, the categories are nominal, meaning they are distinct and lack any inherent hierarchy. For instance, a score of "1 GAP" is as unrelated to a score of "2 Non-GAP" as it is to a score of "3 Do not know". Agreement between annotators occurs only when they select the same category. Consequently, the weight function for nominal data is simply the identity function:

$$w_{kl} = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{if } k \neq 1 \end{cases} \tag{A.3}$$

We use the identity function as the weight function for our data, which is necessary in the next step for calculating p_a.

Step 4: calculating P_a

This represents the most extensive and challenging phase, during which we determine the overall observed percent agreement P_a using the following equation:

$$P_a = P'_a \left(1 - \frac{1}{n\overline{r}_{\cdot}}\right) + \frac{1}{n\overline{r}_{\cdot}} \tag{A.4}$$

Where:

- P': the average value of the observed agreement for each English word i across all the words,
- n: the number of English words,
- $\overline{r}_{\cdot}$: the mean number of ratings per word, and
- $n\overline{r}_{\cdot}$: the total number of ratings.

We describe the application of this equation as shown in the following phases:

1. Calculate $\overline{r}_{\cdot}$, which is the average number of annotators who rated each word. This value was calculated in the step of the building agreement table, $\overline{r}_{\cdot}$ equals 4 as shown in Table A.3.

2. Using Equation (A.5), we calculate $\overline{r}_{ik}$, which is the weighted count of how many annotators gave each word i a rating that fully or partially matched category k.

$$\overline{r}_{ik} = \sum_{l=1}^{q} W_{kl} r_{il} \tag{A.5}$$

Where:

- W_{kl} is the identity function (as described in Appendix A.2),
- q is the number of categories (e.g., in our case, there are 3 categories), and
- r_{il} is the number of annotators that rated the word i with category l.

For our data, Table A.4 shows the weighted count of how many annotators gave word i matching category k. For example, the $\overline{r}_{ik}$ for word 2 (grandchild) and category 1 (GAP) is:

$$\begin{aligned}\overline{r}_{21} &= \sum_{l=1}^{3} W_{1l} r_{2l} \\ &= 1 * 1 + 0 * 2 + 0 * 1 \\ &= 1\end{aligned}$$

Word(i)	Weighted Count ($\bar{r}_{ik}$)		
	GAP	Non-GAP	Do not know
cousin	4	0	0
grandchild	1	2	1
great grandchild	4	0	0
grandfather, gramps, granddad, grandad, granddaddy, grandpa	0	4	0
granddaughter	1	3	0

Table A.4: Weighted count table

3. In this phase, we calculate $P_{a\setminus i}$, the percentage agreement for each word i. Which is equal to the summation of the percentage agreement for all q categories, as shown in Equation (A.6).

$$P_{a\setminus i} = \sum_{k=1}^{q} P_{a\setminus i,k} \tag{A.6}$$

Where $P_{a\setminus i,k}$ is the annotator's percent agreement for a single word and a single category

This value is calculated by Equation (A.7), which normalizes the weighted rating count (computed in Phase 2) by the average number of ratings per word:

$$P_{a\setminus i,k} = \frac{r_{ik}(\overline{r}_{ik} - 1)}{\overline{r}(r_i - 1)} \tag{A.7}$$

For the same example, word 2 (grandchild) and category 1 (GAP), the $P_{a\setminus 2,1}$ is:

$$\begin{aligned}P_{a\setminus 2,1} &= \frac{r_{2,1}(\overline{r}_{2,1} - 1)}{\overline{r}(r_2 - 1)} \\ &= \frac{1 * (1 - 1)}{4 * (4 - 1)} \\ &= 0\end{aligned}$$

Table A.5 displays the annotators' percent agreement for each single word and single category $P_{a\backslash i,k}$, also, it shows percent agreement for each single word across all categories $P_{a\backslash i}$, as shown in the last column in this table. For instance, $P_{a\backslash 1}$, we can just add up the percent agreement for the three categories, as follows:

$$\begin{aligned} P_{a\backslash 1} &= 0 + 0.17 + 0 \\ &= 0.17 \end{aligned}$$

Word(i)	**Percent agreement for a single word and a single category** $P_{(a\backslash i,k)}$			**Percent agreement for a single word** $P_{(a\backslash i,\textbf{all categories})}$
	GAP	**Non-GAP**	**Do not know**	**All categories**
cousin	1	0	0	1
grandchild	0	0.17	0	0.17
great grandchild	1	0	0	1
grandfather, gramps, granddad, grandad, granddaddy, grandpa	0	1	0	1
granddaughter	0	0.5	0	0.5
Total				3.67

Table A.5: Percent agreement table

4. Once we find $P_{a\backslash i}$ for all words, it is time to calculate their average P'_a across these words, which is needed to apply Equation (A.1). We use Equation (A.8) to find P'_a.

$$\begin{aligned} P'_a &= \frac{1}{n}\sum_{i=1}^{n} P_{a\backslash i} \\ &= \frac{3.67}{5} \\ &= 0.73 \end{aligned} \tag{A.8}$$

   Where n is the number of words.

5. Applying Equation (A.4) to find the overall observed percent agreement P_α

$$\begin{aligned} P_a &= P'_a\left(1 - \frac{1}{n\bar{r}_{\cdot}}\right) + \frac{1}{n\bar{r}_{\cdot}} \\ &= 0.73 * \left(1 - \frac{1}{5 * 4}\right) + \frac{1}{5 * 4} \\ &= 0.75 \end{aligned}$$

Step 5: calculating P_e

We calculate the percent agreement that the annotators would achieve by guessing randomly, as shown in Equation (A.9):

$$P_e = \sum_{k,l}^{q} W_{k,l}\pi_k\pi_l \tag{A.9}$$

Here, π_k represents the percentage of total ratings that fall into each rating category k.

Using Table A.3, we calculate π_k for each category, and the results are as follows:

$$\pi_{GAP} = 0.5 \text{ (10 out of 20)}$$
$$\pi_{non-GAP} = 0.45$$
$$\pi_{do\_not\_know} = 0.05$$

If all the annotators were assigning ratings randomly, the probability of one annotator answering k and a second annotator answering l would just be $\pi_k * \pi_l$. We say that the annotators agree when category k at least partially matches category l — when the weight W_{kl} is greater than 0.

Since we are using categorical weights (using the identity function), w_{kl} is 0 unless the categories are an exact match, so $\pi_k = \pi_l$, and we find the overall percent agreement expected by chance by just calculate the sum of π_k^2:

$$\begin{aligned} P_e &= 0.5^2 + 0.45^2 + 0.05^2 \\ &= 0.46 \end{aligned}$$

Step 6: calculating alpha (α)

Finally, we calculate α using Equation (A.1).

$$\begin{aligned} \alpha &= \frac{P_\alpha - P_e}{1 - P_e} \\ &= \frac{0.75 - 0.46}{1 - 0.46} \\ &= 0.54 \end{aligned}$$

For the six steps described above, we implemented a customized Python script to execute these steps and calculate Alpha. The script takes the annotators' responses and outputs the Alpha value. These responses are provided as a multidimensional array, where each row corresponds to an annotator and each column represents the investigated source words. Thus, a cell entry (r_{ij}) denotes the response of annotator i for source word j. See

the Python code implemented for the pilot kinship experiment shown in Figure A.1.

```
pip install krippendorff

from krippendorff import alpha
# example data: list of lists where each sublist
#               represents annotators' responses

responses = [
    ['G', 'N', 'G', 'N', 'N'],
    ['G', 'N', 'G', 'N', 'N'],
    ['G', 'D', 'G', 'N', 'N'],
    ['G', 'G', 'G', 'N', 'G']
]

# Calculate Krippendorff's Alpha
alpha_value = alpha(reliability_data=responses,
                    level_of_measurement='nominal')

print("Krippendorff's Alpha:", alpha_value)
```

Figure A.1: The Python code calculates Alpha for the pilot kinship experiment