UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Building language-independent culture-aware multilingual lexical resources

Nandu Chandran Nair

Advisor

Prof. Fausto Giunchiglia

Università degli Studi di Trento

April 2022

Abstract

Language is an essential part of any society to thrive. Lexical resources are the building blocks of any language; they allow us to find similarities and diversities when comparing languages. However, numerous limitations like funding or lack of expert support hinder language resource development, and consequently, many minor languages are becoming extinct. A possible way to preserve a language is by connecting the lexical resources with famous languages like English. However, the reference language might influence the language development and mapping process. This thesis suggests a methodology for language development and mapping to avoid the supremacy of a reference language. Hence, the thesis presents a strategy to conserve languages to combat one language's dominance over another in the resource. The methodology proposed builds improved and up-to-date concept-oriented multilingual lexical resources from existing ones. The advantage of having such resources is that we can use them to compare the languages, study the differences and similarities, and exploit the information to measure and improve the quality of the languages. Similarly, this thesis shows the importance of the structural organization of multilingual resources to represent the meaning across languages. This thesis focuses on Indian languages, but the methodologies explained are adaptable to be used for any other language. The main outcomes of this thesis are (i) a methodology to create a multilingual resource that does not depend on a reference language and (ii) to present a good quality concept-oriented resource for various Indian languages for the community to preserve the culture.

Keywords

[Lexical resources, Multilingual resources, Diversity-aware, Concept-centered, Indian languages]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Language is an essential part of any society to thrive. It enables people to communicate and express themselves. We can preserve a group's cultural values and knowledge when a language thrives. However, when a language disappears, the knowledge and capacity to understand the language's culture are compromised because teachings, customs, oral traditions, and other inherited knowledge do not transfer to speakers. Furthermore, future generations lose a vital part of the culture that allows them to understand it completely, making language a vulnerable aspect of cultural heritage. Henceforth, it becomes crucial to preserve the languages.

Indian traditions have a wide range of concepts and ideas for personality development. The most comprehensive sources for understanding personality development in traditional Indian philosophy are the Vedas and Upanishads [40]. India has many different civilizations connected to a deep spirituality by the numerous great souls, saints, and yogis. The glorified spirits have revealed many significant facts with their true wisdom. India attracts visitors from all over the world and from many religious backgrounds to experience them.

Indian languages are rich with concepts represented in lexical resources, prioritizing western languages. That means concepts from Indian languages are translated into the concepts available in western languages. For example, the Malayalam word “ഹസ്തസൂത്രം” (hasthasoothram).” The term represents a type of ornament that only married women wear. This word cannot be found in western lexical resources since this Indian word is not that popular. Hence, the word will not give correct meanings unless it is documented. Single words bring entire scenes to mind, and translations will never quite do that justice. Because of this, cultures that lose languages can become isolated, left without a way to express themselves fully. In this context, having a resource that covers in-depth the Malayalam language would help us understand spiritual teachings better. This thesis aims to preserve languages by proposing a methodology to develop lingual resources.

This thesis provides a methodological approach toward a big vision of preserving the languages through lingual resources that equally give importance to all languages in the resource. The thesis highlights the importance of diversity-aware, which takes each language will be treated uniquely as possible, without any supremacy of any language. It is widely acknowledged that languages grow alongside the people who speak them; thus, our methodology considers continuous development rather than just producing a resource. The proposed approach aims to collect the language elements and connect them with world languages to reduce the limitations of knowledge transfer. This thesis also shows that the multilingual resource structure matters to represent and learn the diversity across languages.

1.1 Motivation

1.1.1 Language Preservation

Language is essential to express thoughts and feelings and is a medium to intercommunicate with other humans. When a language is not spoken daily, it tends to disappear; similarly, the rich culture, teaching, values, and customs die with it [36]. Henceforth, it is not passed down from generation to generation. Globalization is the primary cause of the disappearance of many more minor languages. Most native languages are replaced by popularly spoken languages such as English and Spanish. The probability of a language dying increased tenfold because of globalization. Thus, preserving languages is crucial since it is fundamental to preserving entire cultures [43].

1.1.2 Progressive Development

There are around 7000 languages available in the world. Each research group's vision is to collect all languages. However, some languages are rich in their development and maintenance. The progressive development does not depend on the importance of the language [12]. For example, Malayalam is a classical language from India and has around 38 million speakers (based on statistics in 2019). Even with the massive number of speakers, the language is highly unrepresented globally [47]. For example, IndoWordNet's total lexical elements in Malayalam WordNet are approximately 30K synsets, including the named entities and repeated entries. This thesis is motivated to improve the quantity and quality of any language resource progressively and hopes to set up a community for that.

1.1.3 Limit the Digital Language Divide

The popularity of languages was addressed in part before. One factor that helps the popularity of a language is the availability of language on the Internet [26]. Natural language is present on the Internet in various ways, including content, scripts, and the availability of NLP tools [24]. The absence of languages will cause reduced popularity and a linguistic gap in the digital world [29].

To understand this scenario better, let us search for the word “Restaurant” using Google Maps (refer to Figure 1.1). Figure 1.1b shows the search result in English, and Figure 1.1a offers the search in Malayalam. Figure 1.1b satisfies the query even without the translation. However, for Malayalam, Google Maps directed restaurants located in South India. This scenario suggests that Google Maps will respond to local language only in the local area, not globally. This illustration demonstrates the risk posed by the digital language divide. Google Maps may refuse service to a community group or require them to utilize a particular language to receive better service.

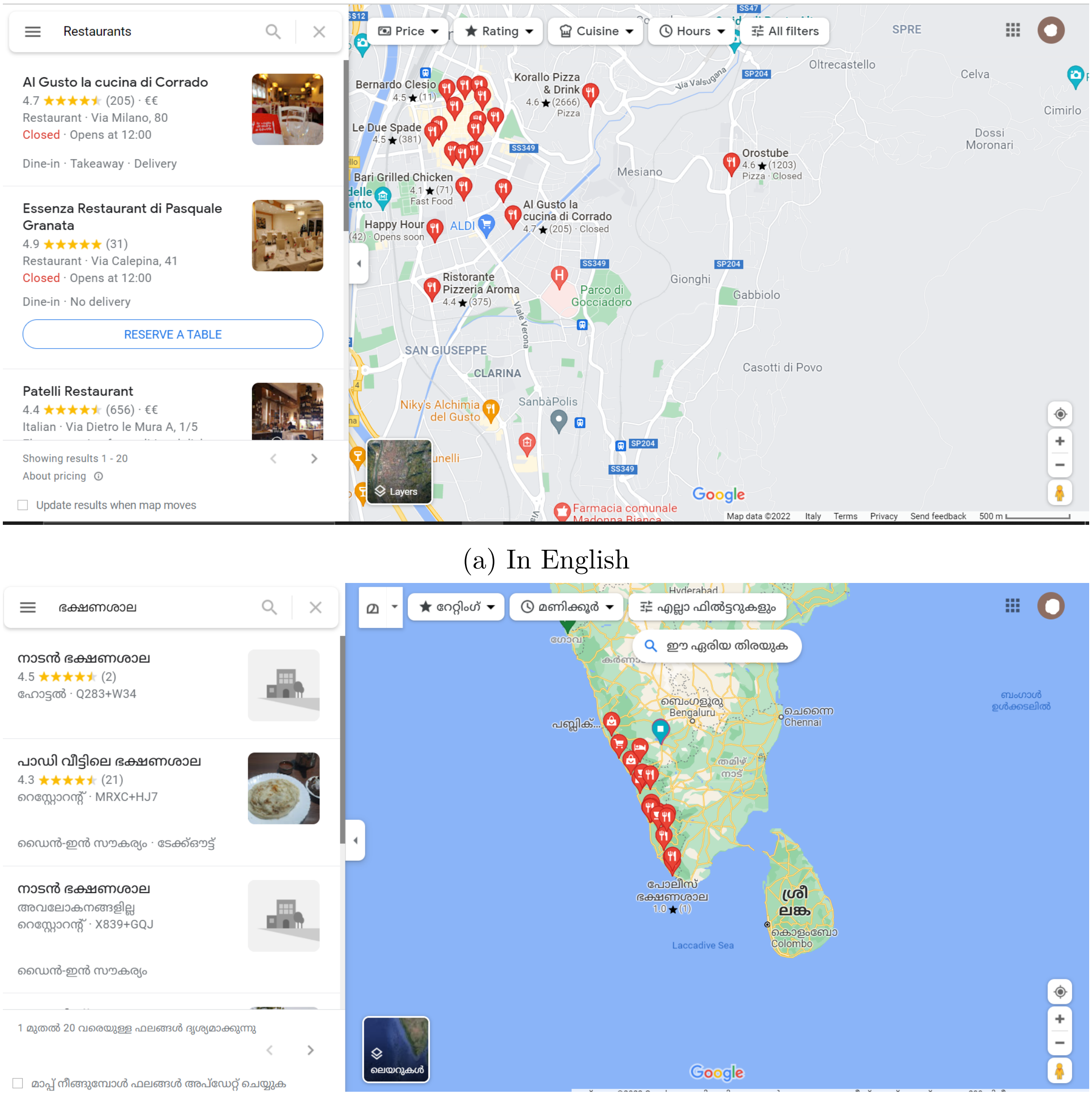
Developing the language resources for more languages and evolving existing language resources helps reduce the digital language divide. This thesis hopes to make languages communicate with other languages without any boundaries and make them famous.

1.1.4 Resource with Good Coverage

Table 1.1: Number of synsets per language

No. of synsets	**No. of languages**
Greater than 100000	2
Between 10000 and 100000	31
Between 1000 and 10000	50
Between 100 and 1000	74
Less than 100	190

The language resources are available in multiple forms. Text corpora [39], speech corpora [25], and terminology databases [28], are a few examples of language resources-this thesis focus on developing a lexico-semantic resource similar to WordNet [30] for Indian languages. WordNet organizes



Restaurants

Al Gusto la cucina di Corrado
4.7 ★★★★★ (205) · €€
Restaurant · Via Milano, 80
Closed · Opens at 12:00
Dine-in · Takeaway · Delivery

Essenza Restaurant di Pasquale Granata
4.9 ★★★★★ (31)
Restaurant · Via Calepina, 41
Closed · Opens at 12:00
Dine-in · No delivery
RESERVE A TABLE

Patelli Restaurant
4.4 ★★★★★ (656) · €€
Italian · Via Dietro le Mura A, 1/5

Showing results 1 - 20
About pricing
□ Update results when map moves

(a) In English

ഭക്ഷണശാല

നാടൻ ഭക്ഷണശാല
4.5 ★★★★★ (2)
ഹോട്ടൽ · Q283+W34

പാഡി വീട്ടിലെ ഭക്ഷണശാല
4.3 ★★★★★ (21)
റെസ്റ്റോറന്റ് · MRXC+HJ7
ഡൈൻ-ഇൻ സൗകര്യം · ടേക്ക്ഔട്ട്

നാടൻ ഭക്ഷണശാല
അവലോകനങ്ങളില്ല
റെസ്റ്റോറന്റ് · X839+GQJ
ഡൈൻ-ഇൻ സൗകര്യം

1 മുതൽ 20 വരെയുള്ള ഫലങ്ങൾ ദൃശ്യമാക്കുന്നു
□ മാപ്പ് നീങ്ങുമ്പോൾ ഫലങ്ങൾ അപ്ഡേറ്റ് ചെയ്യുക

(b) In Malayalam

Figure 1.1: Google Maps search result for query "Restaurant"

the languages in terms of synsets. Synsets represent a set of synonym sets.

Table 1.1 shows the coverage of synsets in languages from a multilingual lexicon-semantic resource of around 300 languages. Here we are assuming that the coverage of the resource is directly proportional to the quality of the resource. As we can see, there are languages with very little data. Even though many resources are there, they are only available to a select group of people. So, both linguists and researchers have to limit their work or need to develop resources from scratch [10]. This thesis will resolve the abovementioned issues and provide the resource to the community of interest to use it.

1.2 Research Questions

We hope to find the solutions to the following research question in this thesis.

1. How to develop a language resource with diversity-aware?

   The thesis considers diversity-aware as embracing the uniqueness of language resources. Multilingual resources are developed mainly by taking a reference language. The reference language helps to connect the common concepts. However, eventually, the resulting resource influences reference language and make it less diversity-aware. The first question this thesis attempts to answer is, “How can a multilingual lexical resource be created without linguistic influence?”

2. How to increase the availability of the concepts from less popular languages in other world languages?

   It is common to know concepts like “pizza” from Italy and “naan” from India worldwide. However, these concepts are limited to specific domains like “food.” We tend to extend the concepts’ popularity to other disciplines.

3. How to represent the diversity of a language in a multilingual lexical resource?

   There are many multilingual resources available. This thesis will compare the significant resources to understand if they represent diversity across languages.

1.3 Proposed Approach

The work focuses on supporting language development from a computer science perspective. The proposed steps are listed below:

1. Methodology for developing a multilingual resource that is diversity-sensitive to preserve languages

   In the approach, we transform the multilingual resource for Indian languages, IndoWordNet (IWN), using the Universal Knowledge Core (UKC) principle, a large-scale multilingual resource [16]. In UKC, different languages codify the meaning by clustering them into concepts. Concepts are linked by semantic relations, forming a network or hierarchy of concepts that express lexical meaning shared across languages. The supra-lingual layer makes the resource more open and easily extensible.

2. Generate IndoUKC, a concept-centered resource for Indian languages

   This thesis introduces a new multilingual lexical resource comprising Indian languages: IndoUKC. IndoUKC represents the meanings of words across different languages using concepts. A language-independent layer in IndoUKC connects the synsets of all Indian languages to the relevant ones from English and other languages. We publish the resource and set it up to enrich it with more language concepts.

1.4 Contributions

Major contributions of this thesis are,

1. a process for developing a multilingual resource independent of a reference language
2. a resource for the community to use in multiple Indian languages, high-quality and concept-oriented

Contributions to the thesis are also described (in part) in the following publications:

- Representing Interlingual Meaning in Lexical Databases –This was submitted in the IJCAI survey track in 2022.
- Is this Enough?-Evaluation of Malayalam WordNet –This was published in the First workshop for Dravidian languages, DravidianLangTech 2021. Refer to chapter 8 for the details of this publication.
- Aligning the IndoWordNet with the Princeton WordNet - This was published in ICNLSP 2019. Chapter 9 details this publication.
- IndoUKC: A Concept Centered Indian Multilingual Lexical Resource –This was published in LREC 2022. Chapter 10 provides the details of this publication.

1.5 Structure of the Thesis

This thesis is structured into six parts: lexical resources, problem, solution, evaluation, implementation, and finally, conclusion and future work. Part I presents a brief review of the literature in the various fields connected with the work presented in this thesis. Chapter 2 discusses the Princeton WordNet, and chapter 3 details multilingual lexical resources. Chapter 4 explains the structure and details of the IndoWordNet, and chapter 5 presents the Universal Knowledge Core. Part II motivates the problem by explaining real-life examples from language technologies that show the issues of a poor-quality multilingual lexical resource. The solution methodology is detailed in Part III: a four-phase process of resource development. The evaluation of the proposed resource in terms of quality metric incompleteness and semantic similarity is covered in Part IV. The implementation of the proposed resource, IndoUKC, is presented in Part V. Part VI wraps up the thesis by presenting the work summary, lessons learned, and possible future works.

Part I

Lexical resources

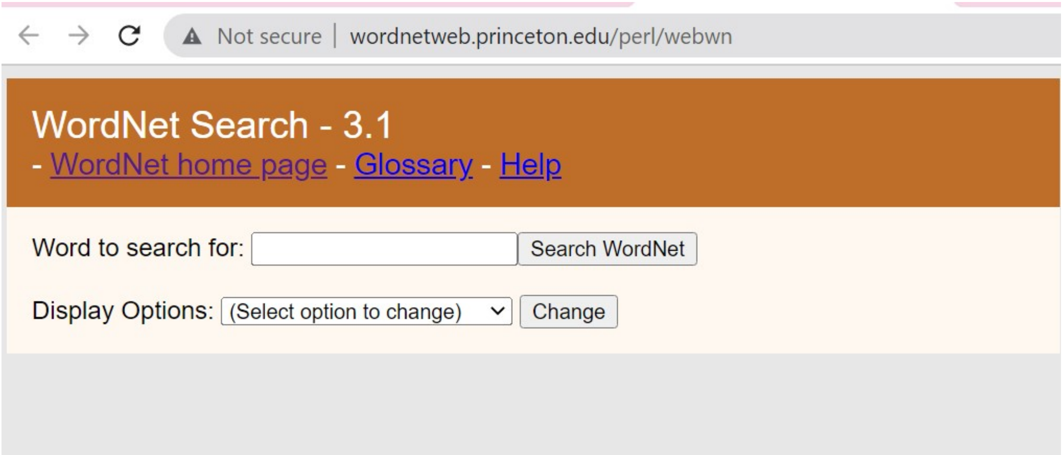
Chapter 2

Princeton WordNet

2.1 Introduction

Princeton WordNet (PWN) was created in the 1980s to understand better how children learn new words. PWN [13] is a commonly used digital language resource because it arranges a lexicon into a set of similar concepts known as synsets and links them to each other. Occasionally users of PWN refer to a lexical ontology because it incorporates some of the ontological relations. PWN provides a single unique beginner, labeled entity. Because ontologies tend to focus on higher-level concepts, a mapping to a lexical resource is advantageous because it extends the ontology's concepts to the leaves of the hierarchies. PWN and the Suggested Upper Merged Ontology (SUMO) are closely related [14].

PWN has become a tool widely used by the Natural Language Processing (NLP) community for applications including information retrieval, reasoning and inferencing, question answering, and machine translation, which often involve reasoning with the meanings of words. Most multilingual lexical resources consider PWN as the foundation for their resources. Making PWN part of the development helped replicate the semantic relations between the synsets. As shown in Figure 2.1, the resulting network of semantically related words and concepts can be accessed using the browser.



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for:

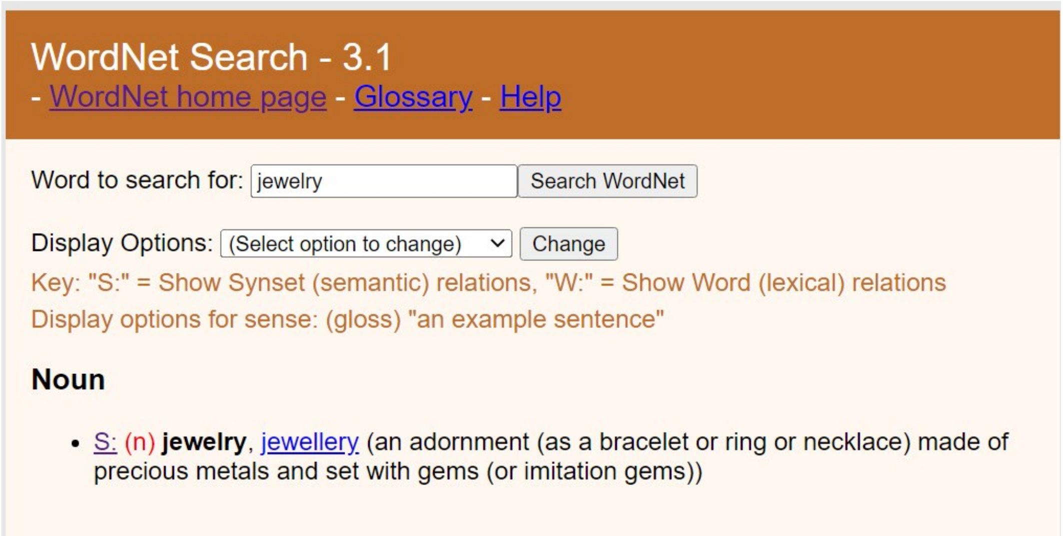
Display Options: (Select option to change)

Figure 2.1: Princeton WordNet Browser

PWN can also be downloaded free by anyone for free.

2.2 Structure and Relations

- **Synsets**: Synsets denote the same concept; they are interchangeable in many contexts and are grouped into unordered sets.



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: jewelry

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

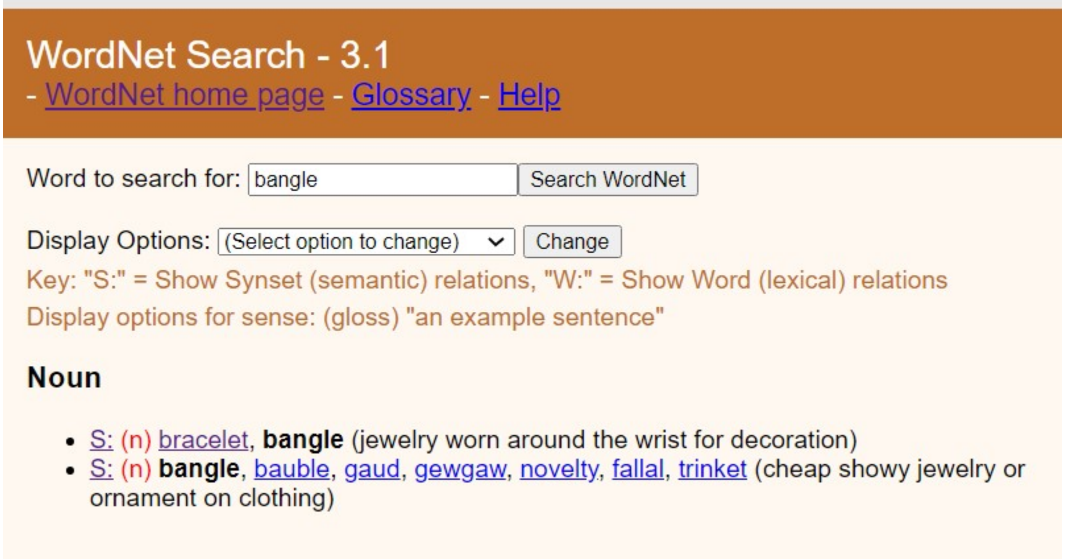
Noun

- S: (n) **jewelry**, jewellery (an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems))

Figure 2.2: Example of synset in PWN

For example, Figure 2.2 shows the example of a synset for the concept of "**jewelry**." It shows that part of speech of the concept is a noun with the label "**n**," and two words in the list have the same meaning and can be used interchangeably.

- **Gloss**: Gloss is the definition of the synset.



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bangle [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

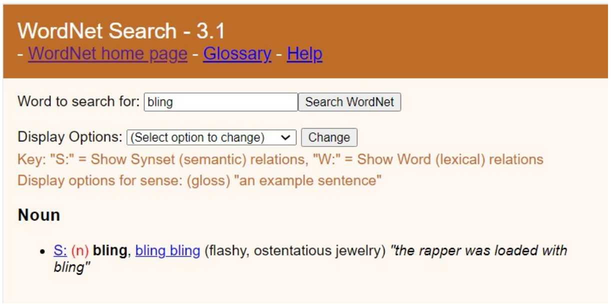
- S: (n) bracelet, **bangle** (jewelry worn around the wrist for decoration)
- S: (n) **bangle**, bauble, gaud, gewgaw, novelty, fallal, trinket (cheap showy jewelry or ornament on clothing)

Figure 2.3: Example of gloss in PWN

For example, Figure 2.3 shows the example of the gloss for the concept "**bangle**." Gloss is represented in brackets for each synset. As shown, the "**bangle**" has two different word forms for the same part of speech, "**noun**" based on the meaning. The first-word form has the gloss, "**jewelry worn around the wrist for decoration**," and the second form has the gloss, "**cheap showy jewelry or ornament on clothing**."

- **Example**: Usually, one or more short sentences illustrate the synset members' use.

An example is not a mandatory field for PWN. However, this field


WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: bling [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bling**, bling bling (flashy, ostentatious jewelry) *"the rapper was loaded with bling"*

Figure 2.4: Example of example in PWN

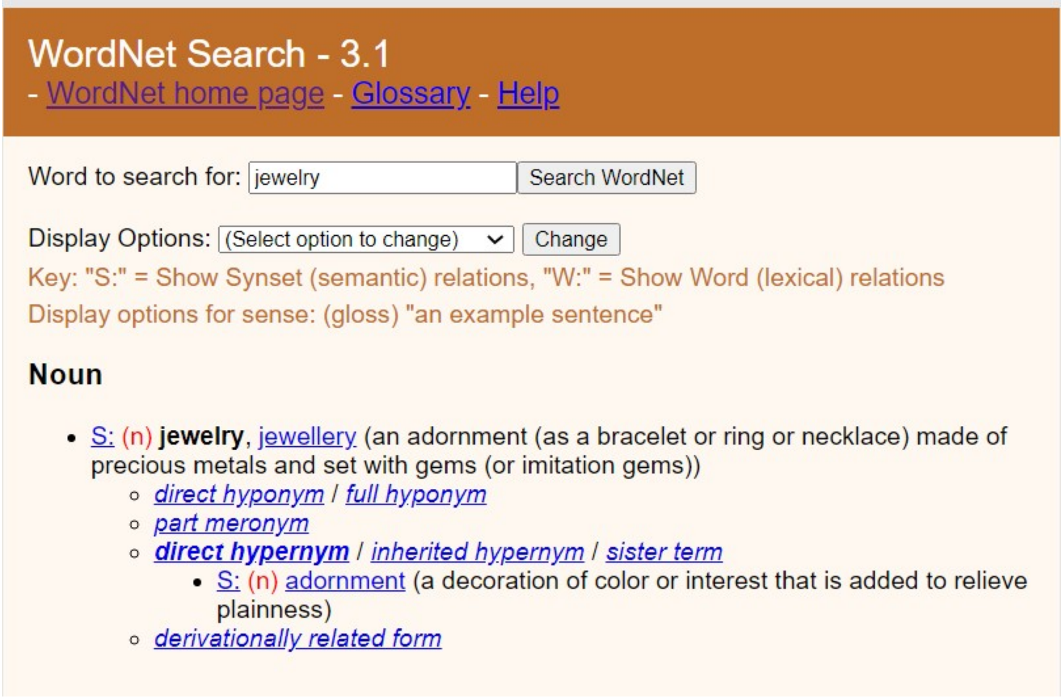
helps to understand the usage. For example, in Figure 2.4, the concept "**bling**" is given. An example sentence is represented in bold in the double quotes. For the concept of "**bling**," the example sentence is "**the rapper was loaded with bling**."

- **Hypernymy, Hyponymy, or ISA relation**: The semantic relationship between a more general and specialized word is known as hypernymy. Hypernymy relation links more general synsets to increasingly specific ones, whereas hyponymy links specific synsets to a general synset.

  PWN will give a direct hypernym or direct hyponym of the selected concept. Figure 2.5 shows the direct hypernym of the concept "**jewelry**," which implies that "**jewelry**" is a hyponym of the concept "**adornment**" or, in other words, the concept "**adornment**" is the hypernym of the concept "**jewelry**." The root node "**entity**" is at the top of all noun hierarchies. Hyponymy relation is transitive.

- **Meronymy**: Meronymy is the part-whole relation held between synsets. Parts are inherited from their superiors and are not inherited "upward" because they may be unique to certain things rather than the entire class.

  For example, Figure 2.6 shows the example of meronymy for the



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: jewelry Search WordNet

Display Options: (Select option to change) Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

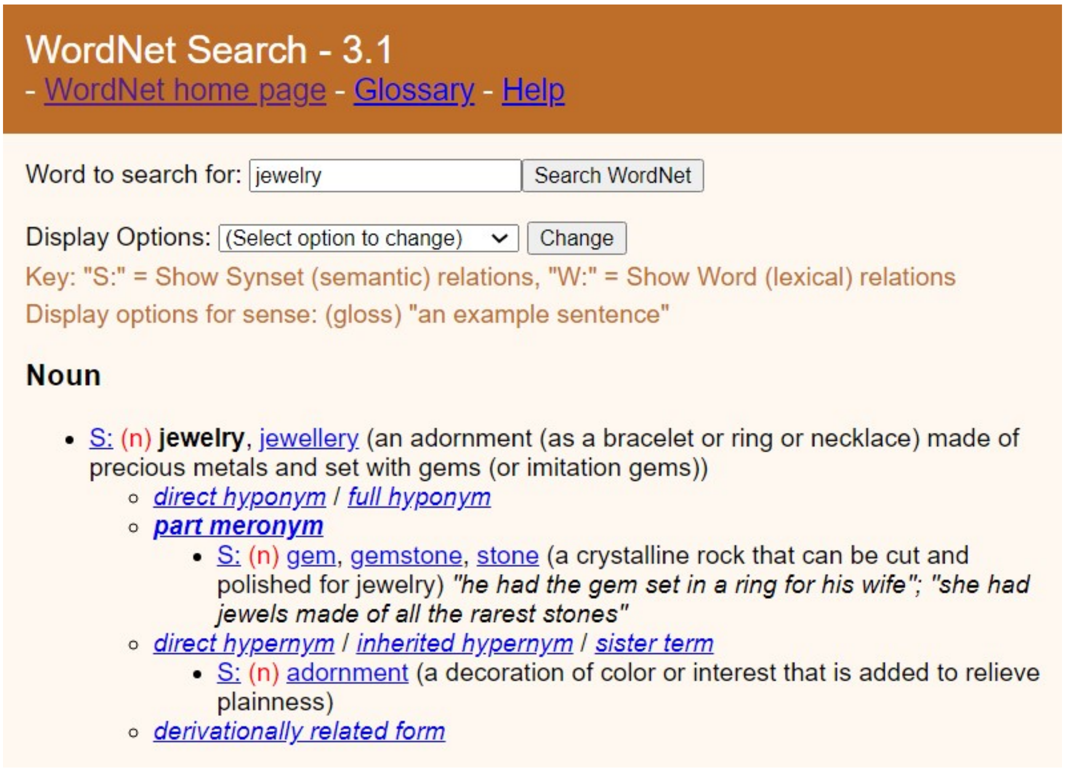
Noun

- S: (n) **jewelry**, jewellery (an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems))
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) adornment (a decoration of color or interest that is added to relieve plainness)
  - *derivationally related form*

Figure 2.5: Example of hyperonymy in PWN

concept of “**jewelry**.” PWN labeled the meronymy relation as part “**meronym**” label, and the synsets “**gem**,” “**gemstone**,” and “**stone**” are meronyms for the concept of “**jewelry**.”

Verb synsets are arranged into hierarchies; verbs towards the bottom of the trees (troponyms) express increasingly specific manners characterizing an event. The specific manner expressed depends on the semantic field; volume is just one dimension along which verbs can be elaborated. Others are speed or intensity of emotion. Verbs that describe occurrences that are inextricably and unidirectionally related are linked. Adjectives are organized in terms of antonymy. Pairs of “direct antonyms” reflect the strong semantic contract of their members. Each polar adjective is linked to several “semantically similar” ones. Semantically similar adjectives are “indirect antonyms” of the central member of the opposite pole. Relational adjectives (“pertainyms”) point to the nouns derived. There are only a few



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: jewelry [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **jewelry**, jewellery (an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems))
    - *direct hyponym* / *full hyponym*
    - ***part meronym***
        - S: (n) gem, gemstone, stone (a crystalline rock that can be cut and polished for jewelry) *"he had the gem set in a ring for his wife"; "she had jewels made of all the rarest stones"*
    - *direct hypernym* / *inherited hypernym* / *sister term*
        - S: (n) adornment (a decoration of color or interest that is added to relieve plainness)
    - *derivationally related form*

Figure 2.6: Example of meronymy in PWN

adverbs in PWN, as most English adverbs are straightforwardly derived from adjectives via morphological affixation.

2.3 Role of PWN in the development of Language Technologies

PWN is a crucial tool for NLP systems that, for example, require lexical disambiguation. PWN is accessible to human users using a web browser to analyze lexical structure and patterns. One of the key technologies for many other NLP applications is word sense discrimination, and semantic relations in wordnet can be utilized for this purpose ([8] shared task for SemEval, reported on by [34]). Using wordnets to increase the precision of information retrieval in a semantic search engine, [1] describe the value of wordnets in more detail. [42] described employing wordnets to autonomously produce vocabulary exams for second language acquisition in relation to language learning applications [6].

Chapter 3

Multilingual Lexical Resources

3.1 Introduction

According to Ethnologue, around seven thousand languages are actively spoken in the world today [7]. Plus, there are all the ancient languages and dialects. All these languages have immense economic and cultural value. However, so far, most of the work on the development of lexical resources has focused on a small number of languages, essentially those spoken in the richest cultures [21]. Beyond single-language lexical resources, multilingual lexical resources have a pivotal role in language technologies such as cross-lingual word sense disambiguation, machine translation, or bilingual lexicon induction [18]. They are crucial for under-resourced languages, complement corpus-based approaches, and link these languages with the rest of the world.

PWN has also motivated the development of many multilingual lexical resources. The most popular multilingual wordnets include EuroWordNet, and Open Multilingual WordNet, which are the following as described.

3.2 EuroWordNet

EuroWordNet (EWN) [46] is a multilingual resource with wordnets of eight European languages: Dutch, Italian, Spanish, English, French, Estonian, German and Czech, with all languages having a hierarchy similar to PWN. The first four wordnets have around 30K synsets and 50K word meanings. EWN used PWN to connect the languages since PWN had extensive coverage and was freely available. EWN generated a central repository of an unstructured list of meanings (from PWN) called Inter Lingual Index (ILI) [45]. The gathered meanings are represented using synsets, and similar to PWN, a gloss is associated with every synset to define its meaning. ILI links word meanings across languages by connecting synsets of all languages that share the same or similar meaning. Synsets with the same or closer meaning share the same id.

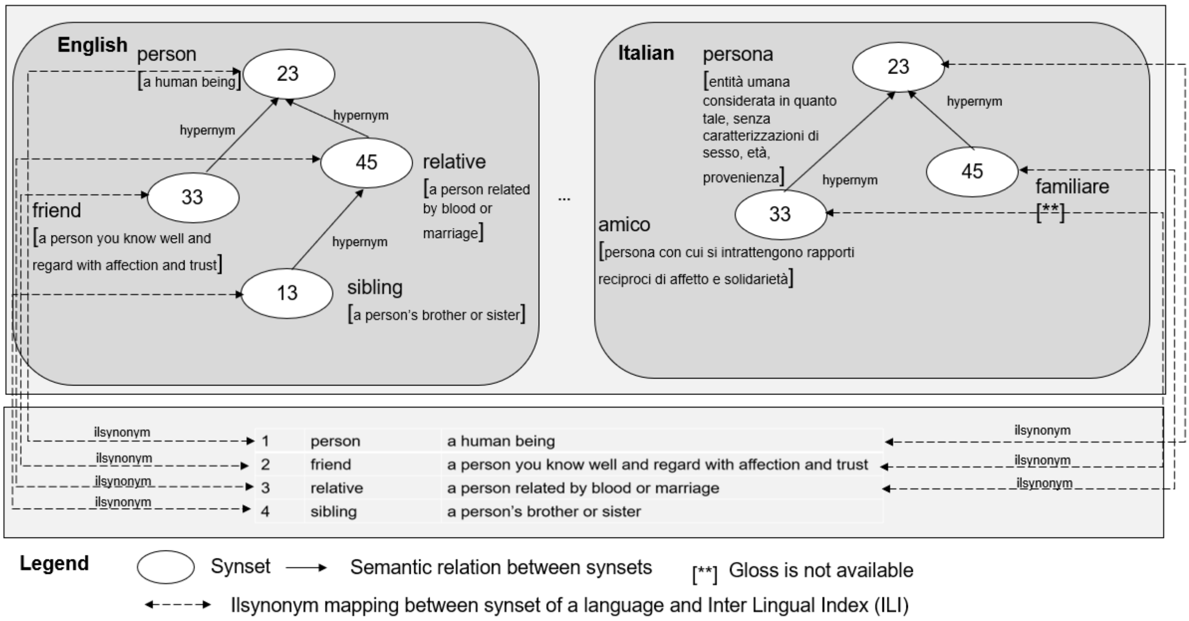

English
person
[a human being]
23
hypernym
hypernym
45
relative
[a person related by blood or marriage]
33
friend
[a person you know well and regard with affection and trust]
hypernym
13
sibling
[a person's brother or sister]
...
Italian
persona
[entità umana considerata in quanto tale, senza caratterizzazioni di sesso, età, provenienza]
23
hypernym
hypernym
45
familiare
[**]
33
amico
[persona con cui si intrattengono rapporti reciproci di affetto e solidarietà]
ilsynonym
1 person a human being
2 friend a person you know well and regard with affection and trust
3 relative a person related by blood or marriage
4 sibling a person's brother or sister
Legend
Synset
Semantic relation between synsets
[**] Gloss is not available
Ilsynonym mapping between synset of a language and Inter Lingual Index (ILI)

Figure 3.1: Interconnection of languages in EuroWordNet

For example, Figure 3.1 shows the diagram for the framework of EWN with two languages that, as can be seen, are composed of two parts of

EWN: the top part shows the languages, and the bottom part shows the list of word meanings in ILI to connect the languages. Each language's synsets are associated with a semantic relation hypernym, and each language has a different hierarchy. The synset **"person"** from English and **"persona"** from Italian was linked to the ILI record **"1: person: a human being"** through the label **IL-SYNONYM**. Also, from Figure 3.1, we could infer that there is no synset for **"sibling"** with the meaning **"a person's brother or sister"** in the Italian language.

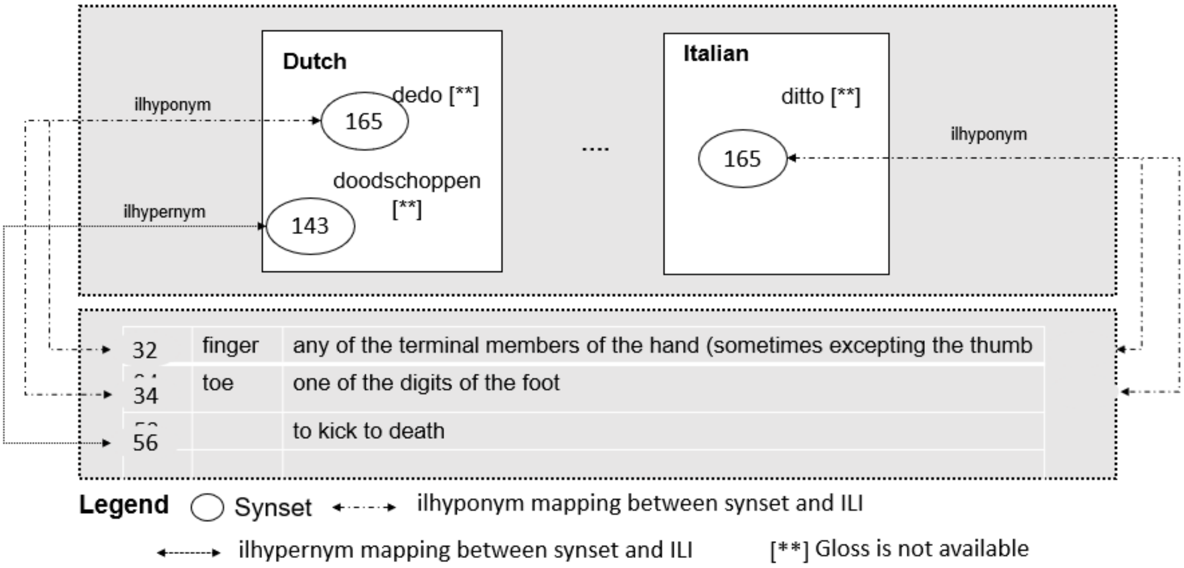

Dutch
dedo [**]
165
doodschoppen [**]
143
ilhyponym
ilhypernym
....
Italian
ditto [**]
165
ilhyponym
32 finger any of the terminal members of the hand (sometimes excepting the thumb
34 toe one of the digits of the foot
56 to kick to death
Legend
Synset
ilhyponym mapping between synset and ILI
ilhypernym mapping between synset and ILI
[**] Gloss is not available

Figure 3.2: Ilhypernym and ilhyponym mapping in EuroWordNet

In Figure 3.2, the synset **"dedo"** from Dutch is connected to the words **"finger"** and **"toe"** in ILI using ilhyponym. Also, the Dutch word "doodschoppen" does not have a corresponding word in PWN; hence, they are connected using ilhypernym. The symbol "**" means the synset's gloss is unavailable.

3.3 Open Multilingual WordNet

The Global WordNet Association (GWA) was established in June 2000 to generate a single platform for many wordnets known as Open Multilingual WordNet (OMW) [5]. OMW uses a centralized language-neutral formal ontology called SUMO to connect the wordnets [35]. The ontology has a multilingual plugin that translates SUMO terms and definitions into many languages. OMW is created around a collection of concepts from the ontology to express the PWN synsets connected to a particular meaning. SUMO definitions and wordnet synsets are used to represent the concepts. For all grid languages, OMW has synsets as a common concept. A synset from a language is mapped to a general concept in SUMO or a concept directly equivalent to the given synsets. The wordnets in the OMW have a semantic network of synsets, but the ontology hierarchy differs from the languages. If the synset of a language is connected to the ontology term, where both have the same meaning, then it is called **equivalent mapping** [14]. If the synset of a language is connected to the Knowledge Interchange Format (KIF) expressions of ontology terms where both have the same meaning, then it is called **subsumption mapping** [14].

The two parts of OMW are depicted in Figure 3.2: the top part shows the languages English, Italian and Spanish, and the bottom part shows the GWG ontology. As opposed to EWN, OMW connects the ontology term to the synset of the languages; therefore, the meaning is mapped differently. The inter-lingual mappings are divided into equivalent mapping and subsumption mapping. For example, the ontology term **“human”** meaning **“a human being”** is mapped to **“persona”** in Italian and Spanish using equivalent mapping. However, the ontology term **“SocialRole”** is connected to synsets **“familiare”** and **“pariente”** through subsumption mapping.

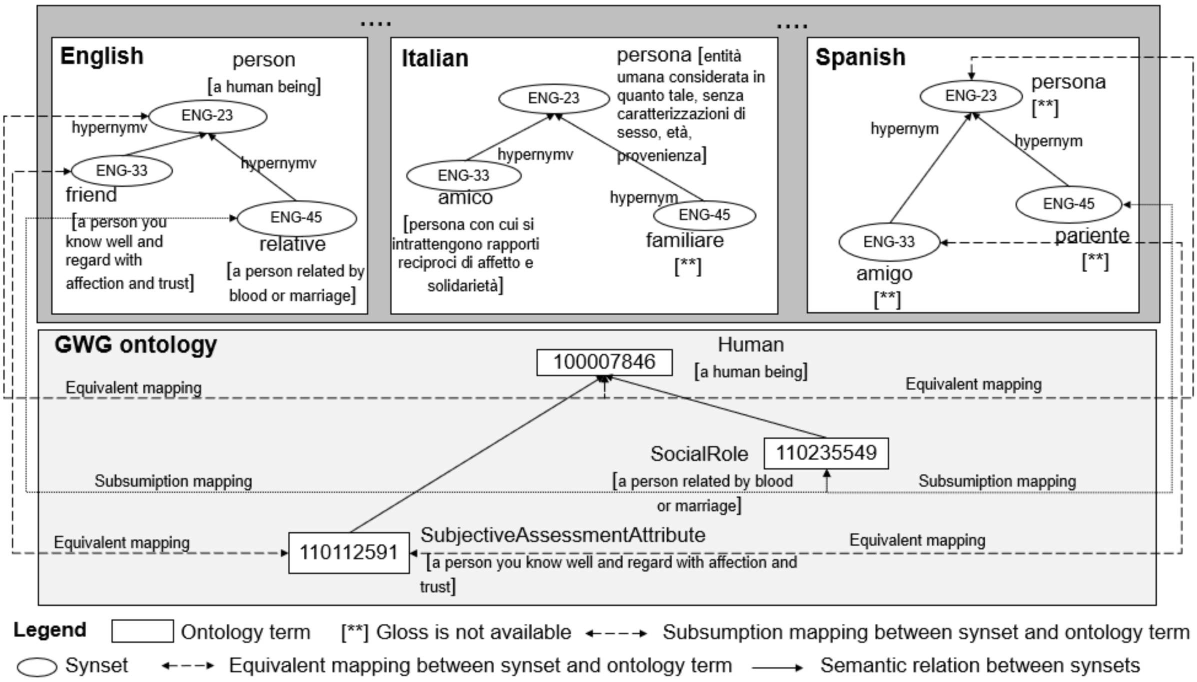

English
person
[a human being]
ENG-23
hypernymv
ENG-33
friend
[a person you know well and regard with affection and trust]
hypernymv
ENG-45
relative
[a person related by blood or marriage]
Italian
persona [entità umana considerata in quanto tale, senza caratterizzazioni di sesso, età, provenienza]
ENG-23
hypernymv
ENG-33
amico
[persona con cui si intrattengono rapporti reciproci di affetto e solidarietà]
hypernym
ENG-45
familiare
[**]
Spanish
ENG-23
persona
[**]
hypernym
hypernym
ENG-45
pariente
[**]
ENG-33
amigo
[**]
GWG ontology
100007846
Human
[a human being]
Equivalent mapping
Equivalent mapping
SocialRole
110235549
[a person related by blood or marriage]
Subsumption mapping
Subsumption mapping
Equivalent mapping
110112591
SubjectiveAssessmentAttribute
[a person you know well and regard with affection and trust]
Equivalent mapping
Legend
Ontology term
[**] Gloss is not available
Subsumption mapping between synset and ontology term
Synset
Equivalent mapping between synset and ontology term
Semantic relation between synsets

Figure 3.3: Open Multilingual WordNet

Recently, efforts toward the second version of OMW were announced [4]. OMW2 replaces the word-to-PWN synset mapping relations of OMW with synset-to-synset mapping relations towards a Collaborative Interlingual Index (CILI). The CILI is a set (i.e., an unstructured collection) of unique IDs representing word meanings relevant to one or more languages. IDs within the CILI are linked to synsets within wordnets with one-to-one equivalence relations (implemented as owl:sameAs in the Semantic Web representation of the OMW2). The collaboratively built and managed CILI is meant to expand beyond PWN to cover synsets with no English equivalents, thus reducing the English-centeredness of OMW. OMW2 also introduces lexical gaps to distinguish between resource incompleteness and non-existent lexicalization.

Figure 3.3 shows an example of Chinese-to-English mapping in OMW and OMW2 for the concept of **“cousin**.” The Chinese word **CW1** is correctly mapped to the English meaning **ES1 relative, relation**. However, the one PWN synset that means cousin is mapped to the eight Chinese terms that denote cousins. The more precise Chinese terms' meanings are lost, and the mappings create the false impression that they are all synonyms and have the same meaning as their English cousins. The false mapping results in a representation that is both incomplete and wrong. For the eight distinct types of Chinese cousins, OMW2 enables the generation of new IDs within the CILI that can be connected to other languages or represented as lexical gaps. Thus, it is possible to include the eight Chinese meanings in the CILI and expressly highlight their absence from the English language's vocabulary.

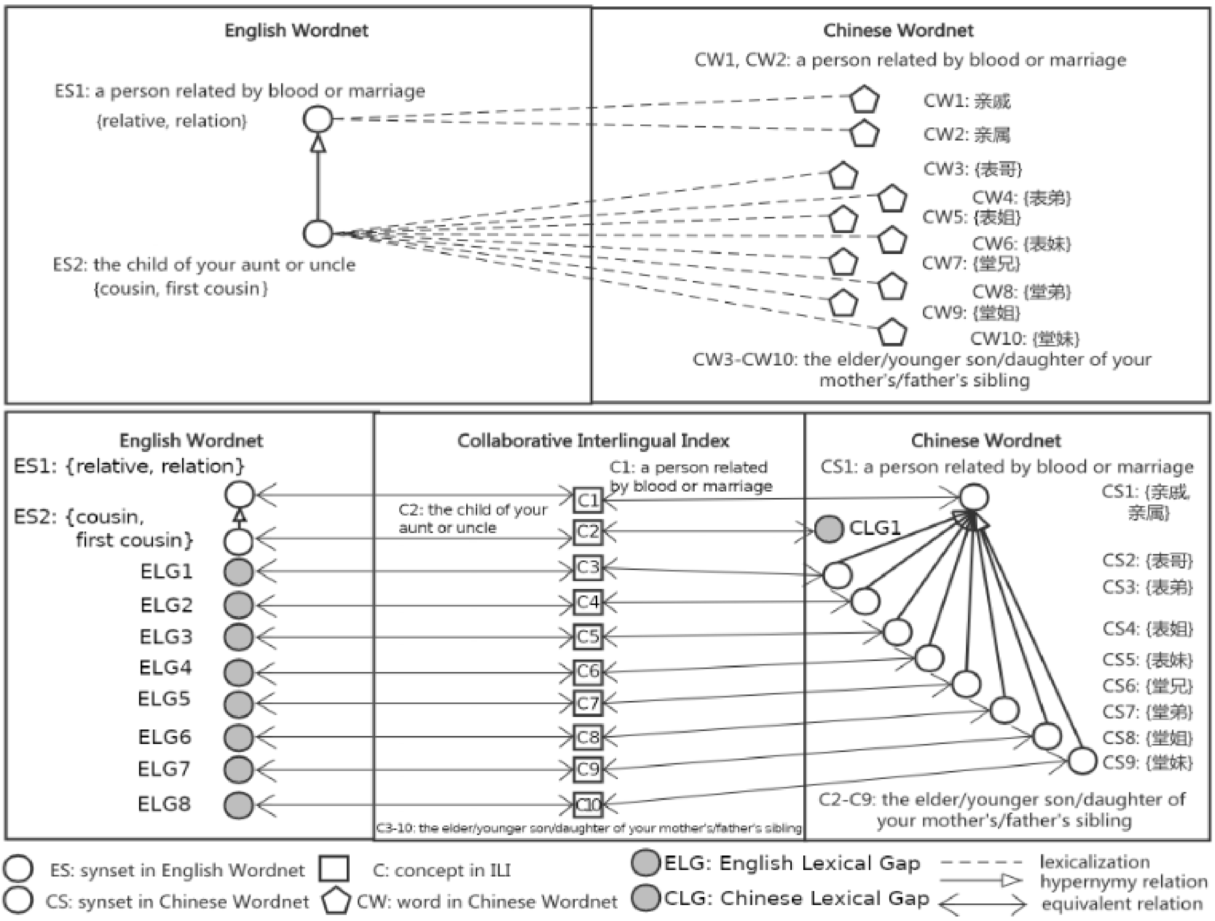

English Wordnet
ES1: a person related by blood or marriage
{relative, relation}
ES2: the child of your aunt or uncle
{cousin, first cousin}
Chinese Wordnet
CW1, CW2: a person related by blood or marriage
CW1: 亲戚
CW2: 亲属
CW3: {表哥}
CW4: {表弟}
CW5: {表姐}
CW6: {表妹}
CW7: {堂兄}
CW8: {堂弟}
CW9: {堂姐}
CW10: {堂妹}
CW3-CW10: the elder/younger son/daughter of your mother's/father's sibling
English Wordnet
ES1: {relative, relation}
ES2: {cousin, first cousin}
ELG1
ELG2
ELG3
ELG4
ELG5
ELG6
ELG7
ELG8
Collaborative Interlingual Index
C1: a person related by blood or marriage
C2: the child of your aunt or uncle
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C3-10: the elder/younger son/daughter of your mother's/father's sibling
Chinese Wordnet
CS1: a person related by blood or marriage
CS1: {亲戚, 亲属}
CLG1
CS2: {表哥}
CS3: {表弟}
CS4: {表姐}
CS5: {表妹}
CS6: {堂兄}
CS7: {堂弟}
CS8: {堂姐}
CS9: {堂妹}
C2-C9: the elder/younger son/daughter of your mother's/father's sibling
ES: synset in English Wordnet
CS: synset in Chinese Wordnet
C: concept in ILI
CW: word in Chinese Wordnet
ELG: English Lexical Gap
CLG: Chinese Lexical Gap
lexicalization
hypernymy relation
equivalent relation

Figure 3.4: Chinese-to-English mappings of meanings using the OMW and OMW2 data models [3]

Let us note that the relationship between **"cousin"** and the eight Chinese meanings is nowhere represented in Figure 3.4. On the one hand, this is understandable as the knowledge that the synset **ES2** is more general than that of **CS2**–**CS9**, so the knowledge is not directly derivable from the two wordnets. On the other hand, even if one wanted to specify such a relation manually, it would not be representable within the framework using the CILI and equivalence mappings alone. As the CILI layer leaves hierarchical structuring of word meanings to individual wordnets, it cannot express cross-lingual hierarchical relationships.

Chapter 4

IndoWordNet

4.1 Introduction

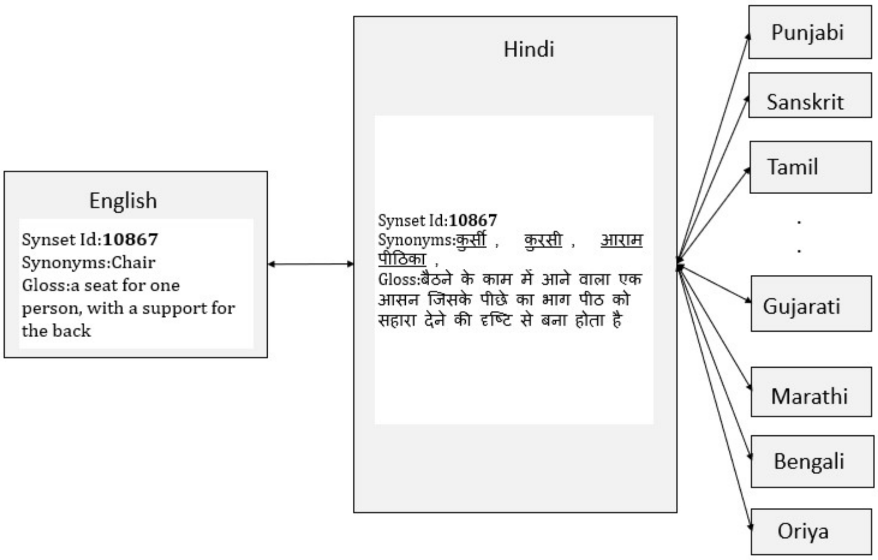

English
Synset Id:10867
Synonyms:Chair
Gloss:a seat for one person, with a support for the back
Hindi
Synset Id:10867
Synonyms:कुर्सी , कुरसी , आराम पीठिका ,
Gloss:बैठने के काम में आने वाला एक आसन जिसके पीछे का भाग पीठ को सहारा देने की दृष्टि से बना होता है
Punjabi
Sanskrit
Tamil
.
.
Gujarati
Marathi
Bengali
Oriya

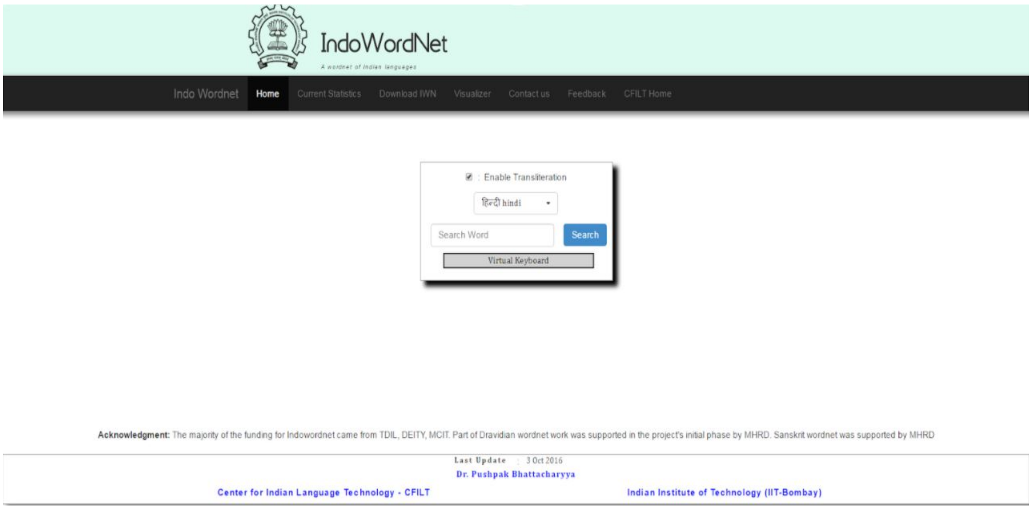
Figure 4.1: Linked IndoWordNet structure

India's languages represent a vital component of the world's linguistic diversity. Four language typologies are spoken on the Indian subcontinent: Indo-European, Dravidian, Tibeto-Burman, and Austroasiatic. According to a list of languages collected by several native speakers, many languages are in the top 10 in the world in terms of the number of people who speak them, including Hindi-Urdu (5th), Bangla (7th), Marathi (12th), and so on. Thus, it is vital from a technological, scientific, and linguistic standpoint to develop wordnets for Indian languages.

IndoWordNet (IWN) is the first multilingual wordnet for Indian languages, developed by the joint efforts of universities across India [11]. Eighteen languages from Indo-Aryan, Dravidian, and Sino-Tibetan language families are included in IWN. Hindi WordNet (HWN), [33] developed by IIT Bombay, India, is used as the central repository to connect other languages [11]. The wordnets are created using an expansion approach from the HWN. The HWN was developed from the first principles and was the first wordnet for an Indian language. The method adopted was the same as the PWN. Figure 4.1 shows the concept of **“chair”** represented in the Hindi language. Hindi is the central node, and other languages are connected. Hence, the hierarchy of the IWN is taken from the HWN.

IWN is highly similar to EWN. However, the pivot language is Hindi which, of course, is linked to the English wordnet or PWN. Also, typical Indian language phenomena like complex predicates and causative verbs are captured in IWN. Like EWN, IWN has a central repository to connect all the languages, and all languages in IWN share identical ids. IWN uses the semantic network of Hindi synsets to connect the languages. Also, the coverage of the synsets in IWN is more than in EWN.

IWN is publicly browsable (see Figure 4.2). The Indian language wordnet building efforts forming the sub-components of the IWN project are the North East WordNet project, the Dravidian WordNet Project, and the Indradhanush project, all funded by the TDIL project.



☑ : Enable Transliteration

हिन्दी hindi

Acknowledgment: The majority of the funding for Indowordnet came from TDIL, DEITY, MCIT. Part of Dravidian wordnet work was supported in the project's initial phase by MHRD. Sanskrit wordnet was supported by MHRD

Last Update : 3 Oct 2016

Dr. Pushpak Bhattacharyya

Center for Indian Language Technology - CFILT

Indian Institute of Technology (IIT-Bombay)

Figure 4.2: IndoWordNet browser

4.2 Hindi WordNet

Researchers from the Center for Indian Language Technology of the Computer Science and Engineering Department of IIT Bombay developed the HWN. The main language of India is Hindi, which belongs to the Indo-Iranian language group. With almost 500 million speakers worldwide, it is the fifth most widely spoken language, according to [22]. Inspired by the well-known wordnet for the English language, HWN is the first wordnet for an Indian language. Through an interface, we can browse the HWN (see Figure 4.3). HWN goes beyond being a simple Hindi dictionary. HWN provides various relationships between synonym sets, or synsets, which stand in for individual concepts.

HWN contains nouns, verbs, adjectives, and adverbs. The following factors make up each entry:

- Synset: a group of interchangeable terms.
- Gloss: the concept. It consists of two parts:
  - Text definition: it defines the concept denoted by the synset.

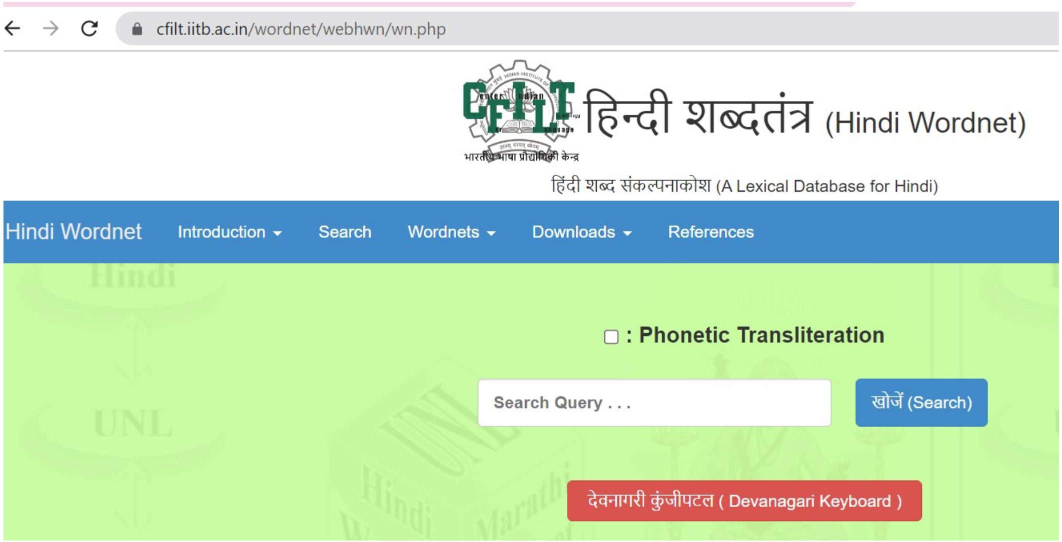
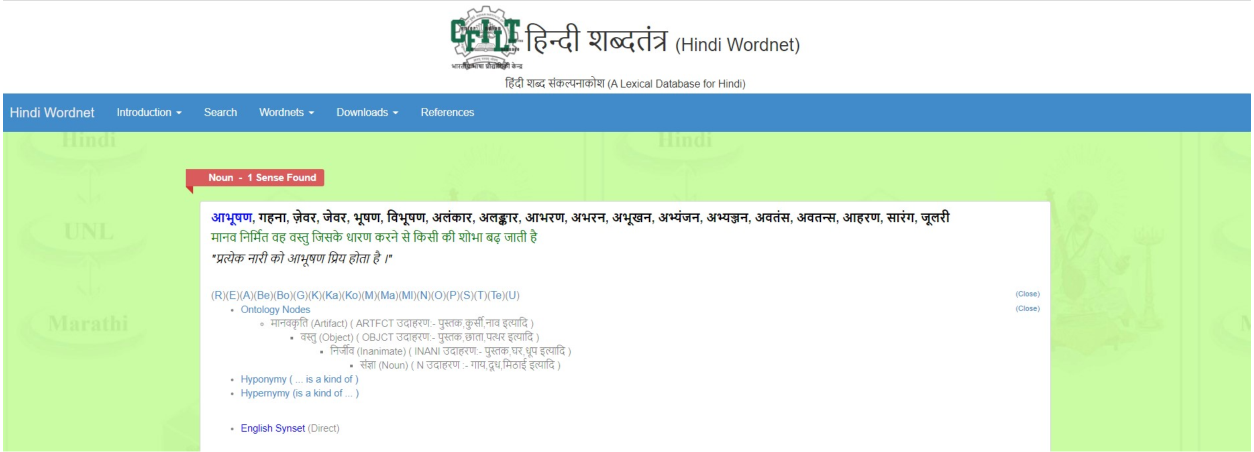

cfilt.iitb.ac.in/wordnet/webhwn/wn.php
CFILT
भारतीय भाषा प्रौद्योगिकी केन्द्र
हिन्दी शब्दतंत्र (Hindi Wordnet)
हिंदी शब्द संकल्पनाकोश (A Lexical Database for Hindi)
Hindi Wordnet
Introduction
Search
Wordnets
Downloads
References
□ : Phonetic Transliteration
Search Query . . .
खोजें (Search)
देवनागरी कुंजीपटल (Devanagari Keyboard)

Figure 4.3: Hindi WordNet browser

  - Example sentence: it explains how the words in the statement are used. In most cases, the words in a synset can be replaced in a sentence.

- Position in Ontology: a hierarchical structure of concepts, or more particularly, a categorization of entities and activities, is what an ontology is. Each syntactic category has its ontological hierarchy (noun, verb, adjective, adverb). Each synset is associated with a node in the ontology.

For example, as shown in Figure 4.4, the example of synsets for the concept **jewelry** are "आभूषण (abhooshan)," "गहना (gahana)," "ज़ेवर(jsevar)," "भूषण (bhooshan)" and so on. The text definition for the synsets is "मानव निर्मित वह वस्तु जिसके धारण करने से किसी की शोभा बढ़ जाती है (manav nirmith vah vasthu jiske dharan karne se kisi ki Shobha badu jathi he)," which translated into English as "man-made thing that enhances one's beauty by wearing" and the corresponding example sentence for the synsets is "प्र–त्येक नारी को आभूषण प्रिय होता है (pretheyk nari ko abhooshan priy hotha he)" which translated into English as "especially, woman loves jewelry" The



हिन्दी शब्दतंत्र (Hindi Wordnet)

हिंदी शब्द संकल्पनाकोश (A Lexical Database for Hindi)

Noun - 1 Sense Found

आभूषण, गहना, ज़ेवर, जेवर, भूषण, विभूषण, अलंकार, अलङ्कार, आभरण, अभरन, अभूखन, अभ्यंजन, अभ्यञ्जन, अवतंस, अवतन्स, आहरण, सारंग, जूलरी

मानव निर्मित वह वस्तु जिसके धारण करने से किसी की शोभा बढ़ जाती है

"प्रत्येक नारी को आभूषण प्रिय होता है ।"

(R)(E)(A)(Be)(Bo)(G)(K)(Ka)(Ko)(M)(Ma)(Ml)(N)(O)(P)(S)(T)(Te)(U) (Close)

- Ontology Nodes (Close)
  - मानवकृति (Artifact) (ARTFCT उदाहरण:- पुस्तक,कुर्सी,नाव इत्यादि)
    - वस्तु (Object) (OBJCT उदाहरण:- पुस्तक,छाता,पत्थर इत्यादि)
      - निर्जीव (Inanimate) (INANI उदाहरण:- पुस्तक,घर,धूप इत्यादि)
        - संज्ञा (Noun) (N उदाहरण :- गाय,दूध,मिठाई इत्यादि)
- Hyponymy (... is a kind of)
- Hypernymy (is a kind of ...)

- English Synset (Direct)

Figure 4.4: Example of entry in Hindi WordNet

ontology nodes of the concept are "artifact," "object," "inanimate," and "noun."

4.3 Malayalam WordNet

Other than HWN, there are other Indian languages in IWN: Assamese, Bengali, Bodo, Gujarati, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. Linkage of HWN with Malayalam and other chosen Indian languages creates a multilingual resource for Indian languages, useful for many NLP applications.

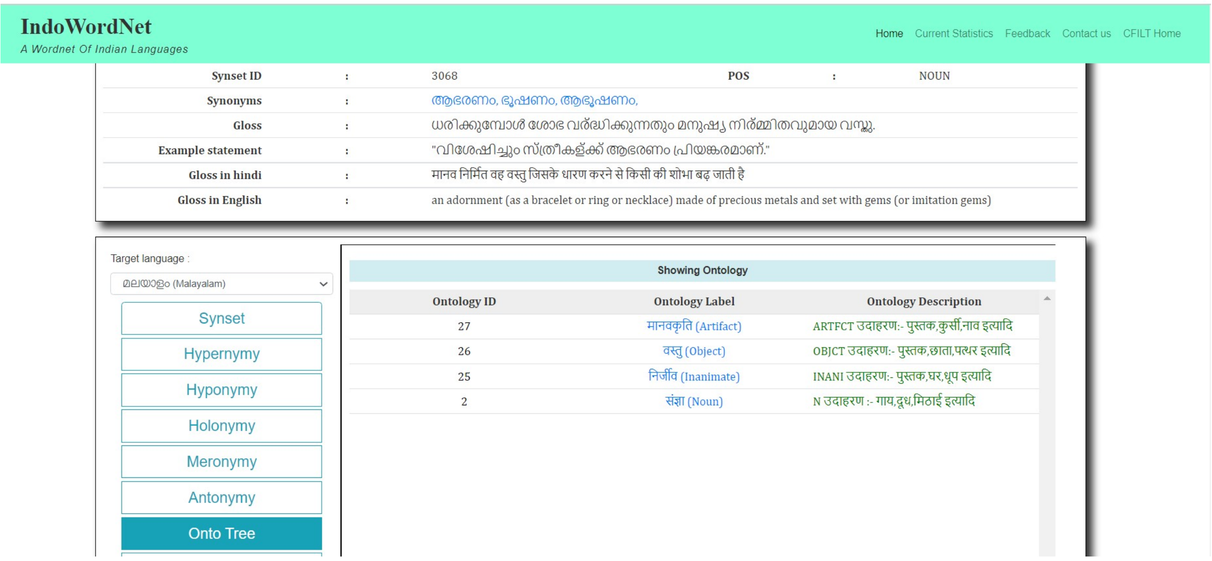
The Indian University, Amrita Vishwa Vidyapeetham, took part in the development of Malayalam WordNet (MWN) [37] in 2011 as part of the project entitled "Development of Dravidian wordnet: an integrated wordnet for Telugu, Tamil, Kannada, and Malayalam," along with other Indian languages, MWN was later incorporated into IWN. MWN is a component of Dravidian wordnet, which is the component of IWN. MWN aims to capture the network of lexical or semantic relations between lexical items or words in Malayalam. The development of MWN is motivated by HWN.

MWN uses the same structure and lists the following entries for each concept: synset ID, synsets, gloss, and example sentence. Synset ids used in MWN differ from PWN as the expand approach was used to develop MWN. In the expanded approach, Hindi synsets are translated into Malayalam; the translation approach ensures concepts that are culturally relevant to India are retained as part of MWN.

MWN was developed using the expansion approach. Synsets are produced using this method by referencing the language's existing wordnet. Malayalam synsets are created using Hindi as a source language. A synset linkage tool provided by the Indian Institute of Technology, Bombay, is used to develop synsets in Malayalam. This synset linking program has a graphical user interface on the left that shows the Hindi synset and an interface on the right that allows you to enter the Malayalam synset. Lexicalization of concepts varies across languages, leading to synsets in one language but not another. The lexical items are divided into six categories:

1. universal,
2. Pan-Indian,
3. in-family,
4. language-specific,
5. rare, and
6. synthesized

The lexical items covered consist of nouns, verbs, adjectives, and adverbs. The project's primary goal is to clarify word meanings. In that sense, marking will be done at the project' s next stage. The sense-making will be done on the corpus using sense IDs as tags. This will allow for disambiguation of word senses in the text. There are 30139 synsets available in MWN; of these, 20071 synsets are nouns, 3311 synsets are verbs, 501 synsets are adverbs, and 6256 synsets are adjectives.



IndoWordNet
A Wordnet Of Indian Languages

Synset ID	:	3068	POS	:	NOUN
Synonyms	:	ആഭരണം, ഭൂഷണം, ആഭൂഷണം,			
Gloss	:	ധരിക്കുമ്പോൾ ശോഭ വർദ്ധിക്കുന്നതും മനുഷ്യ നിർമ്മിതവുമായ വസ്തു.			
Example statement	:	"വിശേഷിച്ചും സ്ത്രീകള്ക്ക് ആഭരണം പ്രിയങ്കരമാണ്."			
Gloss in hindi	:	मानव निर्मित वह वस्तु जिसके धारण करने से किसी की शोभा बढ़ जाती है			
Gloss in English	:	an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems)			

Target language : മലയാളം (Malayalam)

Showing Ontology

Ontology ID	Ontology Label	Ontology Description
27	मानवकृति (Artifact)	ARTFCT उदाहरण:- पुस्तक,कुर्सी,नाव इत्यादि
26	वस्तु (Object)	OBJCT उदाहरण:- पुस्तक,छाता,पत्थर इत्यादि
25	निर्जीव (Inanimate)	INANI उदाहरण:- पुस्तक,घर,धूप इत्यादि
2	संज्ञा (Noun)	N उदाहरण :- गाय,दूध,मिठाई इत्यादि

Figure 4.5: Example of entry in Malayalam WordNet

For example, as shown in Figure 4.5, the example of synsets for the concept of **jewelry** is "ആഭരണം (abharanam)," " ഭൂഷണം (bhooshanam), " and " ആഭൂഷണം (abhooshanam)." The text definition for the synsets is "ധരിക്കുമ്പോള് ശോഭ വര്ദ്ധിക്കുന്നതും മനുഷ്യ നിര്മ്മിതവുമായ വസ്തു (dharikkumbol shobha vardhikkunnathum manushya nirmithavumaya vasthu)" which translated into English as "man-made thing that enhances one's beauty by wearing" and the corresponding example sentence for the synsets is "വിശേഷിച്ചും സ്ത്രീകള്ക്ക് ആഭരണം പ്രിയങ്കരമാണ് (visheshichum sthreekalakku abharanam priyankaramanu)" which translated into English as "especially woman loves jewelry." The ontology nodes of the concept are "artifact," "object," "inanimate," and "noun." The entry also includes the gloss in Hindi and English.

4.4 Interconnection of Languages

A subset of PWN is connected to IWN to support the machine translation applications between English and Hindi, where PWN is taken as the source [9], [38]. Synsets from HWN are connected with other languages' synsets with synonymy or hypernymy relations [41]. The use of HWN (as opposed to PWN) as the hub makes sense for cultural and linguistic proximity to other languages of India (at least concerning to Indo-Aryan languages). Still, the limitation of word meanings to what is lexicalized in Hindi restricts the extent of diversity that the IWN can express.

Figure 4.6 shows the inter-lingual mapping between the synset of Hindi and other Indian languages in IWN: Malayalam and Marathi. The first part shows the Hindi synsets tree mapped with other languages' synsets in the second part. The other Indian languages share the hierarchy of the HWN.

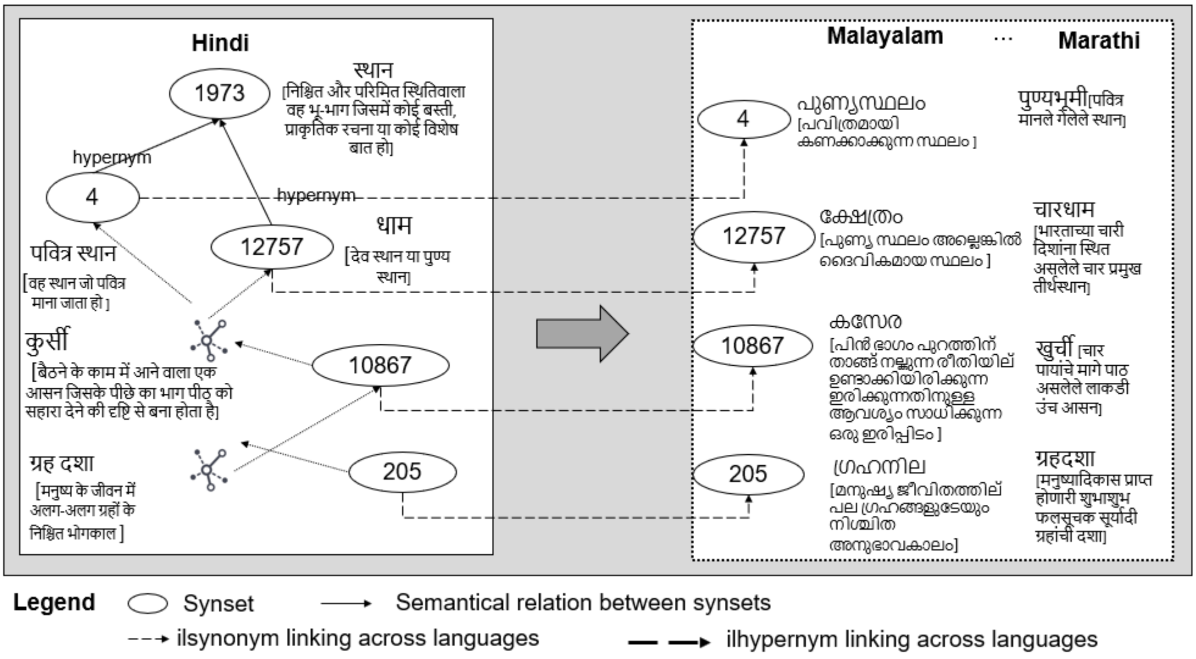

Hindi
1973
स्थान
[निश्चित और परिमित स्थितिवाला वह भू-भाग जिसमें कोई बस्ती, प्राकृतिक रचना या कोई विशेष बात हो]
hypernym
4
पवित्र स्थान
[वह स्थान जो पवित्र माना जाता हो]
hypernym
12757
धाम
[देव स्थान या पुण्य स्थान]
कुर्सी
[बैठने के काम में आने वाला एक आसन जिसके पीछे का भाग पीठ को सहारा देने की दृष्टि से बना होता है]
10867
ग्रह दशा
[मनुष्य के जीवन में अलग-अलग ग्रहों के निश्चित भोगकाल]
205
Malayalam
...
Marathi
4
പുണ്യസ്ഥലം
[പവിത്രമായി കണക്കാക്കുന്ന സ്ഥലം]
पुण्यभूमी[पवित्र मानले गेलेले स्थान]
12757
ക്ഷേത്രം
[പുണ്യ സ്ഥലം അല്ലെങ്കിൽ ദൈവികമായ സ്ഥലം]
चारधाम
[भारताच्या चारी दिशांना स्थित असलेले चार प्रमुख तीर्थस्थान]
10867
കസേര
[പിൻ ഭാഗം പുറത്തിന് താങ്ങ് നല്കുന്ന രീതിയില് ഉണ്ടാക്കിയിരിക്കുന്ന ഇരിക്കുന്നതിനുള്ള ആവശ്യം സാധിക്കുന്ന ഒരു ഇരിപ്പിടം]
खुर्ची [चार पायांचे मागे पाठ असलेले लाकडी उंच आसन]
205
ഗ്രഹനില
[മനുഷ്യ ജീവിതത്തില് പല ഗ്രഹങ്ങളുടേയും നിശ്ചിത അനുഭാവകാലം]
ग्रहदशा
[मनुष्यादिकास प्राप्त होणारी शुभाशुभ फलसूचक सूर्यादी ग्रहांची दशा]
Legend
Synset
Semantical relation between synsets
ilsynonym linking across languages
ilhypernym linking across languages

Figure 4.6: Hindi and other Indian languages wordnets in IndoWordNet

Figure 4.7 shows the inter-lingual mapping between English and Hindi in IWN. The first part shows PWN, and the second part shows the tree of Hindi synsets. The English synset with id 10867 is connected to the synset in Hindi using ilsynonym mapping. While English synsets with id 4 and 12757 have the same meaning and are connected to two synsets with different meanings in Hindi using ilhypernym mapping. Because English synsets are connected to Hindi synsets using either direct linkage or hypernymy linkage [38].

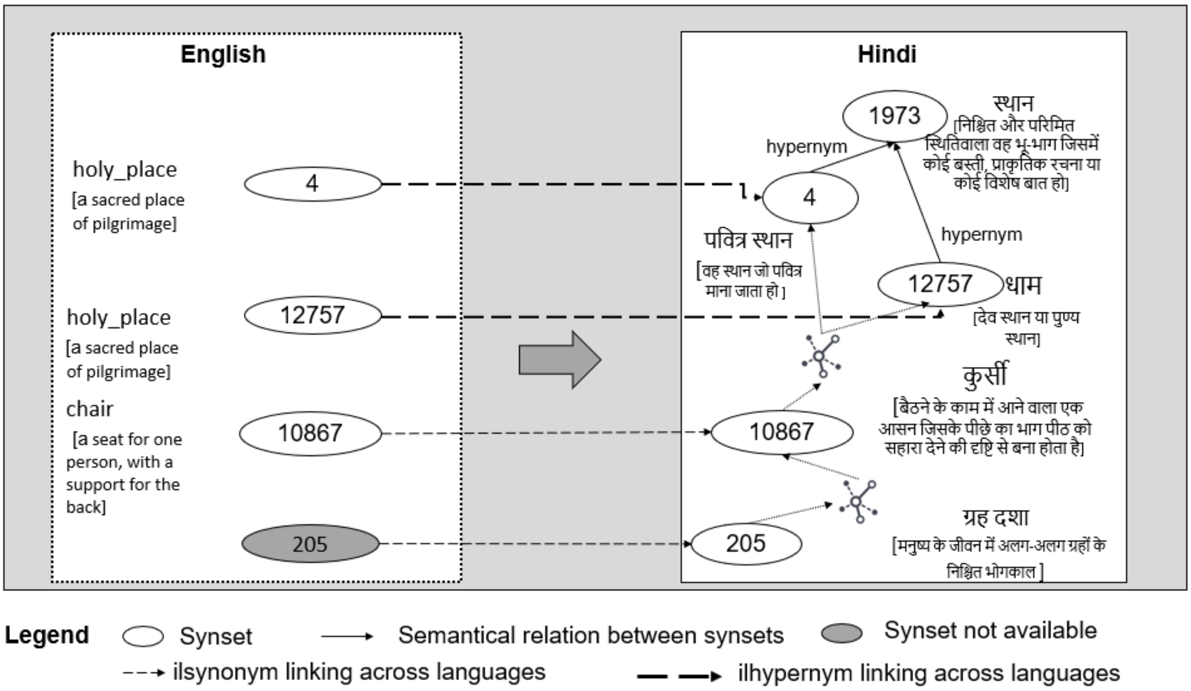

English
holy_place
[a sacred place of pilgrimage]
4
holy_place
[a sacred place of pilgrimage]
12757
chair
[a seat for one person, with a support for the back]
10867
205
Hindi
1973
स्थान
[निश्चित और परिमित स्थितिवाला वह भू-भाग जिसमें कोई बस्ती, प्राकृतिक रचना या कोई विशेष बात हो]
hypernym
4
पवित्र स्थान
[वह स्थान जो पवित्र माना जाता हो]
hypernym
12757
धाम
[देव स्थान या पुण्य स्थान]
कुर्सी
[बैठने के काम में आने वाला एक आसन जिसके पीछे का भाग पीठ को सहारा देने की दृष्टि से बना होता है]
10867
ग्रह दशा
[मनुष्य के जीवन में अलग-अलग ग्रहों के निश्चित भोगकाल]
205
Legend
Synset
Semantical relation between synsets
Synset not available
ilsynonym linking across languages
ilhypernym linking across languages

Figure 4.7: English and Hindi wordnets in IndoWordNet

The IWN is unique because it uses equivalence and hypernymy as cross-lingual relations. Figure 4.8 shows our *cousin's* mappings between Hindi and Malayalam, a Dravidian language from Southern India. In Malayalam, MS1 can be mapped to **HS1** using equivalent mapping, but **MS2**–**MS17** are more specific meanings than **HS2**–**HS9**, which do not exist in HWN.

The solution of IWN is to link them to a more general synset with hypernymy relations: it maps **HS2** (father's sister's son) in Hindi to two more specific Malayalam meanings, **MS2** and **MS3** (father's sister's el-

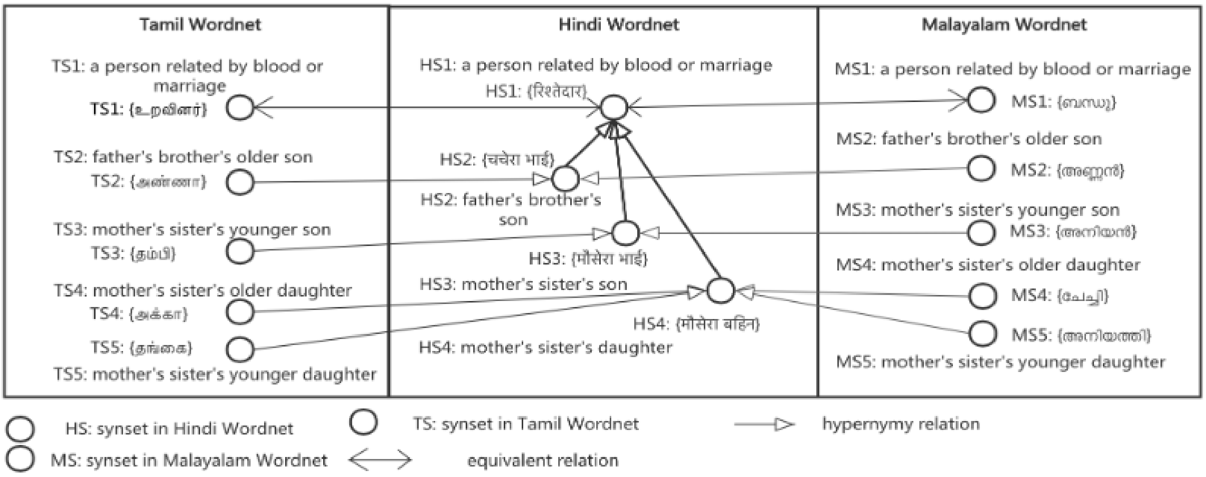

Tamil Wordnet
TS1: a person related by blood or marriage
TS1: {உறவினர்}
TS2: father's brother's older son
TS2: {அண்ணா}
TS3: mother's sister's younger son
TS3: {தம்பி}
TS4: mother's sister's older daughter
TS4: {அக்கா}
TS5: {தங்கை}
TS5: mother's sister's younger daughter
Hindi Wordnet
HS1: a person related by blood or marriage
HS1: {रिश्तेदार}
HS2: {चचेरा भाई}
HS2: father's brother's son
HS3: {मौसेरा भाई}
HS3: mother's sister's son
HS4: {मौसेरा बहिन}
HS4: mother's sister's daughter
Malayalam Wordnet
MS1: a person related by blood or marriage
MS1: {ബന്ധു}
MS2: father's brother's older son
MS2: {അണ്ണൻ}
MS3: mother's sister's younger son
MS3: {അനിയൻ}
MS4: mother's sister's older daughter
MS4: {ചേച്ചി}
MS5: {അനിയത്തി}
MS5: mother's sister's younger daughter
HS: synset in Hindi Wordnet
TS: synset in Tamil Wordnet
hypernymy relation
MS: synset in Malayalam Wordnet
equivalent relation

Figure 4.8: Example of Tamil–Hindi–Malayalam mappings using the IndoWordNet model [3]

der/younger son) through two hypernymy relations, which theoretically is a many-to-one hypernymy mapping. IWN is thus capable of correctly mapping non-equivalent synsets across languages. On the other hand, because Hindi is the hub, IWN cannot map equivalent meanings across Indian languages if the meaning is not part of Hindi. For example, both Tamil and Malayalam have lexicalizations for the *mother's sister's elder daughter* (**TS4** and **MS4**, respectively), but the IWN can only indicate that they are both hyponyms of **HS4**, which results in information loss.

4.5 Summary

The total number of unique concepts from all 18 languages is 40856. Moreover, there are 483757 synsets from these languages. Table 4.1 shows the list of languages based on their range of synsets. Four languages have more than 35K synsets: Oriya, Gujarati, Bengali, and Hindi. Nepali and Assamese languages have fewer synsets than others, with only slightly more than 10K synsets.

Table 4.1: Number of synsets in languages

Languages	Total Synsets	Noun	Adjective	Verb	Adverb
Assamese	14954	9064	3803	1675	412
Bengali	36333	27271	5812	2804	444
Bodo	15781	8786	4287	2294	414
Gujarati	35570	26493	5827	2805	445
Hindi	39239	28962	6145	3266	473
Kannada	22027	12757	5983	3118	169
Kashmiri	29441	21026	5361	2652	399
Konkani	32356	23136	5741	2998	481
Malayalam	29047	19114	6156	3284	493
Manipuri	16313	10152	3804	2019	332
Marathi	29700	21513	4878	2821	487
Nepali	11659	6718	3211	1469	261
Oriya	35275	27211	5270	2417	377
Punjabi	32336	23239	5820	2836	441
Sanskrit	23117	17578	4028	1245	263
Tamil	25417	16311	5827	2802	477
Telugu	21087	12077	5775	2793	442
Urdu	34105	25129	5744	2792	438

Chapter 5

The Universal Knowledge Core

5.1 Introduction

The Universal Knowledge Core (UKC) is a large-scale multilingual resource developed at the University of Trento, Italy. UKC's structure is designed as a multilayered ontology with a language-independent semantic layer called the **Concept Core (CC)** and a language-specific lexico-semantic layer called the **Language Core (LC)**. The principle of UKC is that there is a clear division of languages used to describe the world as it is perceived and what is being described [15]. The CC is the UKC representation of the world, where all the language-independent concepts form a semantic network. The concepts are interconnected using semantic relations. Each concept is identified using a unique id. The LC contains lexicalization of the concepts in different languages, and each synset is associated with one language and a concept from CC. The gloss is associated with the synset. The ids of concepts and synsets are different, and each language has different ids. The LC contains the lexicalization of one or more languages. Each of these languages corresponds to a local language core. The LC organizes the relations between synsets and senses of the same language through lexical and lexical-semantic relations. While lexical-semantic relationships hold between synsets, lexical relations hold between senses. For instance, the lexical relation known as the antonym expresses the semantic contrast between two senses. The lexical-semantic relationships, however, still apply to synsets. When expressing, for example, that two synsets have similar meanings, the relation similar-to is employed.

As a result, the UKC uses synsets and lexical concepts to represent the meanings of words. The UKC defines the lexical concept [17] as the language-independent concepts that group the synsets from different languages with the same meaning. The UKC considers a concept as a representation denoting a set of events rather than a set of instances [16]. For a lexical concept to be created, there must be at least one language where it is lexicalized. In the UKC, the semantic relations link the lexical concepts, not the synsets. Currently, the UKC uses the same semantic relations as the PWN.

5.2 Interconnection of Languages

The UKC can be seen as the central hub of all the Local Knowledge Cores (LKCs), where each LKC has its user interface that facilitates the localization process. As shown in Figure 5.1, each LKC has a knowledge base of an independent language and culture that reflects the unique cultural context, human heritage, and glorious history of that culture.

The source and target languages make up a language pair in an LKC. The localization (synsets of the source language) that is included in the LC in a language like English or any other language in the LC is the source language. The target language is the one in which these synsets are localized. The translation, validation, and approval processes localize the synsets. The synsets produced when these processes are complete will be incorporated into the LC of the UKC. The localization process involves more than just translation. It could add new concepts to the CC and

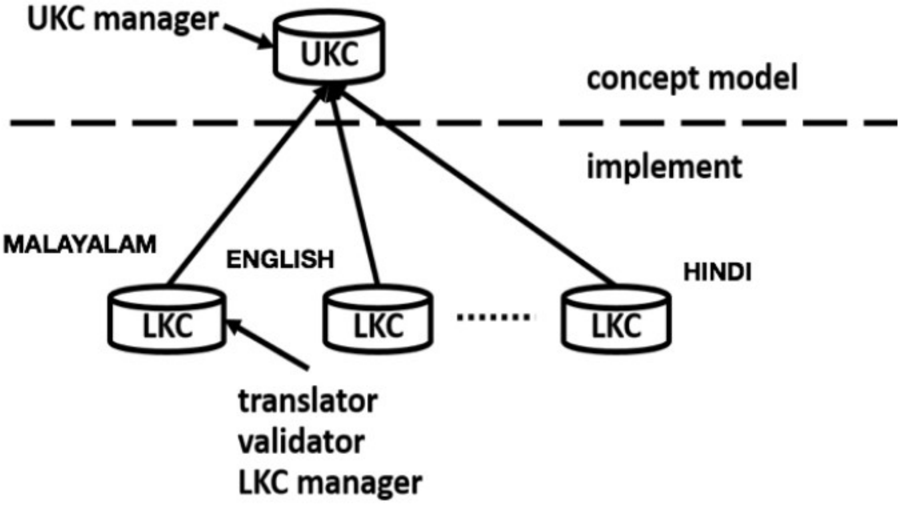


Figure 5.1: Universal Knowledge Core and Local Knowledge Core

identify lexical gaps.

Figure 5.2 shows the concepts linked with synsets from English and two Indian languages, Malayalam and Hindi. LC has the vocabulary for the concepts “chair,” “seat,” and “furniture” in English and Indian languages.

The UKC is one of the most reliable multilingual resources due to its unbiased nature toward language and culture. More importantly, a new language can be easily integrated into this multilingual wordnet by connecting the synsets to the CC [16]. The UKC allows large-scale quantitative studies about language diversity and similarity [15]. The UKC also supports cross-lingual data integration [2].

5.3 Summary

UKC contains about 2.5∼million words in over 1,000 languages, integrating various resources such as PanLex, OMW, IWN, CogNet, and others [3]. There are 1174 languages, 1774815 words, 2648619 language-specific word senses, 110579 concepts, and 14251 lexical gaps [3].

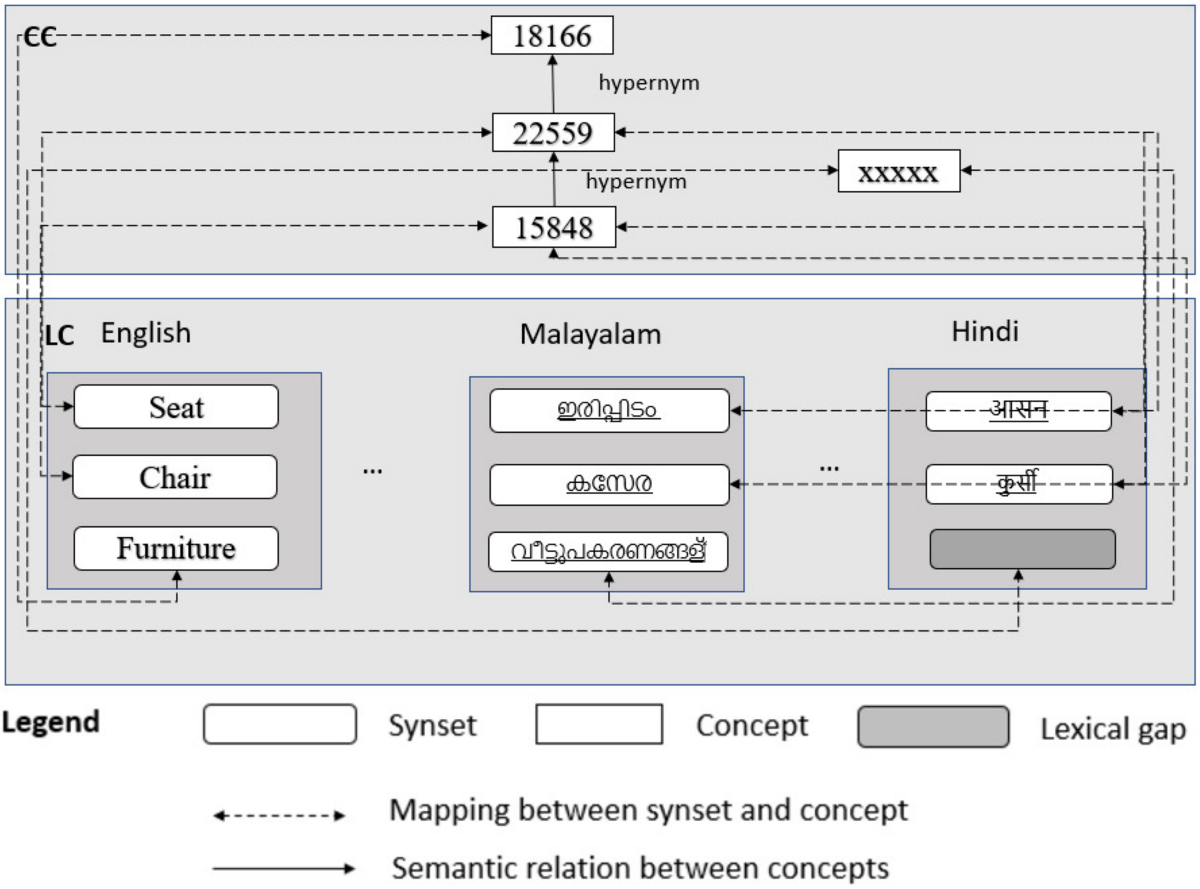


Figure 5.2: Indian languages in Universal Knowledge Core

Part II

Problem

Chapter 6

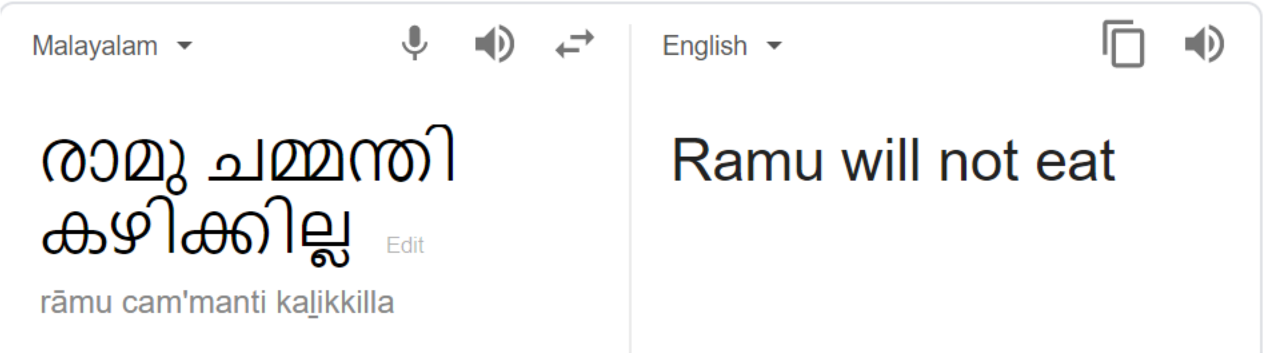
Limitations in preserving Indian Languages with diversity-aware

6.1 Introduction

In today' s multilingual lexical resources, the meanings or concepts of most languages are underrepresented. Such languages, many of which are on the point of extinction and suffer greatly from underrepresentation. This chapter tries to draw attention to the problem of Indian languages not being given the same preference as English. Dominant languages like English accurately depict their lexical meaning spaces, whereas languages with varied linguistic or cultural backgrounds approximate their lexical meaning spaces. In this chapter, we draw attention to structural limitations in the available resources for the Indian language that reduces their expressivity in capturing culturally specific words.

6.2 Diversity-Aware in multilingual lexical resource

Diversity-aware lexical resources ensure that the lexicalization of concepts in one language is not constrained or varied due to that language' s dominance. We employ a reference language to connect the other languages in



രാമു ചമ്മന്തി കഴിക്കില്ല

rāmu cam'manti kaḻikkilla

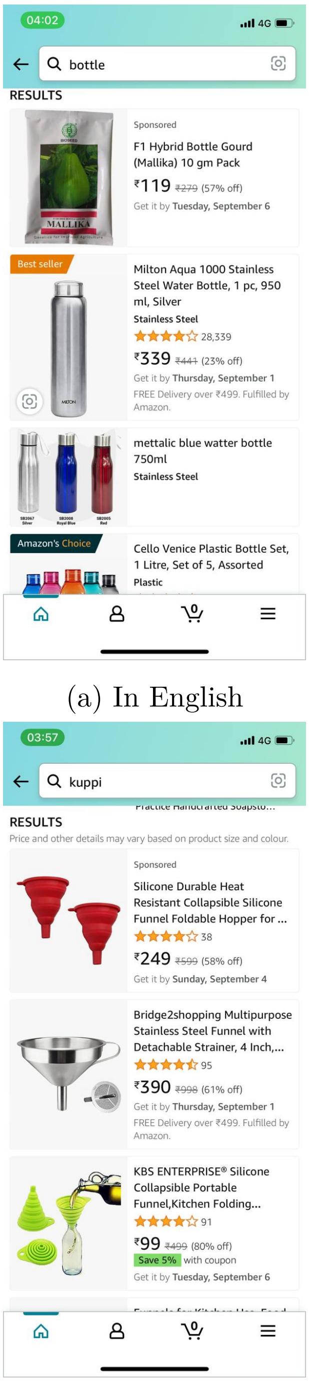
Ramu will not eat

Figure 6.1: Google Translator Example

the multilingual resource. In other words, IWN utilizes Hindi and OMW uses English for linking other languages. To demonstrate our diversity awareness, we provide a cultural example from India. The concept “cousin” is described in PWN as “son or daughter of uncle or aunt.” In Indian culture, this concept does not have a direct lexical value in Malayalam. For example, one concept has a sense of being the “older son of father’ s sister,” and another is the “older son of mother’ s sister.” Linguistically, we could approximate both concepts as close to “cousin” ; however, culturally, both have differences in respect and compassion. As each language is intended to describe different concepts, it is possible to represent them without approximations.

In our thesis, we primarily questioned whether it was appropriate to have one language be dominant. We also questioned whether this would impact how other languages were used for NLP in these resources. We can see that a few terms still need to be recognized in Google Translate for Malayalam, even in 2019. In Figure 6.1, for instance, Google Translate is used to translate a sentence from Malayalam to English. Figure 6.1 shows a sample Malayalam sentence: “രാമു ചമ്മന്തി കഴിക്കില്ല(Ramu chammanthy kazhikkilla),” should be translated as “Ramu will not eat chammanthy.” The machine could not comprehend the word “chammanthy” because it refers to an Indian dish. What the subject "Ram" eats was not specified by the translator. In this case, the translator misidentifies a word in Malayalam that refers to a dish item from Kerala, where Malayalam is spoken. Rich languages like English and Hindi will not have a problem with missing words. Instead of adding every word in a language to provide a complete translation, our goal is to correct or prevent misinterpretation.

Coverage of lexical resources is another limitation. For instance, Amazon India enables searching for products using terms from Indian lexicons rather than English translations. In other words, we can write the word in Malayalam and type it in English to search on Amazon. Nevertheless, resources with these data are needed to better search for concepts peculiar to culture. With an example of a product search on the Amazon India application, we present Figure 6.2. Figure 6.2b displays the results of a search using a Malayalam term written in English. "Kuppi," which means "bottle," is what we look for on Amazon. Figure 6.2a displays the results of a search for the word "bottle" in English. In Figure 6.2a, the search results provided a "funnel" for pouring into the "bottle." The retrieval algorithm interpreted the concept differently as we see the search results differences. For Malayalam, instead of showing results for "bottle," we get the results for the item related to "bottle." Considering the future, we show the importance of having resources with correctly aligned concepts between languages. In this context, at least, it is not a sensitive situation, but in future interpreting, the concept for the local language needs to be done. Such kind of situation occurs due to improper alignment between languages. The possibility of improper alignments is that the Malayalam concept does not have a direct translation, so it will align with a parent concept. Another possibility is that the lexical items of Malayalam languages have not been completed as other languages align. In IWN, the Malayalam language has almost 30K synsets, even though it is one of India's ancient



04:02

bottle

RESULTS

Sponsored

F1 Hybrid Bottle Gourd (Mallika) 10 gm Pack

₹119 ~~₹279~~ (57% off)

Get it by **Tuesday, September 6**

Best seller

Milton Aqua 1000 Stainless Steel Water Bottle, 1 pc, 950 ml, Silver

Stainless Steel

★★★★☆ 28,339

₹339 ~~₹441~~ (23% off)

Get it by **Thursday, September 1**

FREE Delivery over ₹499. Fulfilled by Amazon.

mettalic blue watter bottle 750ml

Stainless Steel

Amazon's Choice

Cello Venice Plastic Bottle Set, 1 Litre, Set of 5, Assorted

Plastic

(a) In English

03:57

kuppi

Practice Handcrafted Soapsto...

RESULTS

Price and other details may vary based on product size and colour.

Sponsored

Silicone Durable Heat Resistant Collapsible Silicone Funnel Foldable Hopper for ...

★★★★☆ 38

₹249 ~~₹599~~ (58% off)

Get it by **Sunday, September 4**

Bridge2shopping Multipurpose Stainless Steel Funnel with Detachable Strainer, 4 Inch,...

★★★★☆ 95

₹390 ~~₹998~~ (61% off)

Get it by **Thursday, September 1**

FREE Delivery over ₹499. Fulfilled by Amazon.

KBS ENTERPRISE® Silicone Collapsible Portable Funnel,Kitchen Folding...

★★★★☆ 91

₹99 ~~₹499~~ (80% off)

Save 5% with coupon

Get it by **Tuesday, September 6**

Funnels for Kitchen Use, Food

(b) In Malayalam

Figure 6.2: Search in Amazon India

and classical languages.

6.3 Open Multilingual Wordnet

Open Multilingual WordNet (OMW) includes Indian languages, although they are in a different repository created by the IWN team. As we already established, IWN is an independent project, and the IWN team used its schema structure. The OMW team's integration with OMW is hampered as a result. In addition, Hindi is employed in IWN instead of English as a reference language. Therefore, resources created by PWN can be combined more quickly. Development without PWN ultimately impedes the growth of under-resourced languages.

Additionally, it will be difficult to map a new concept in an Indian language with no equivalent in PWN. Furthermore, there is no updated version of the material available. If IWN is continued as an independent program, Indian languages will interact less with other world languages.

6.4 IndoWordNet

Two methods are used in IWN to connect the separate wordnets: one involves translating Hindi lexical items into the target language's lexical value, and the other involves aligning the Hindi lexical item with the target language's already created lexical item. The first strategy relies only on Hindi and is less likely to introduce new vocabulary into the target language. When the source and target languages, like Hindi and Marathi, are from close language families or share a common cultural heritage, conceptual alignment between the two languages is successful. Aligning Malayalam and Hindi, however, takes more effort. For instance, "मौसेरा बाई (mousera bhai)" is the son of your father's sister. However, there is no direct translation for Malayalam because it takes into account whether the son of the father's sister is older or younger than the person.

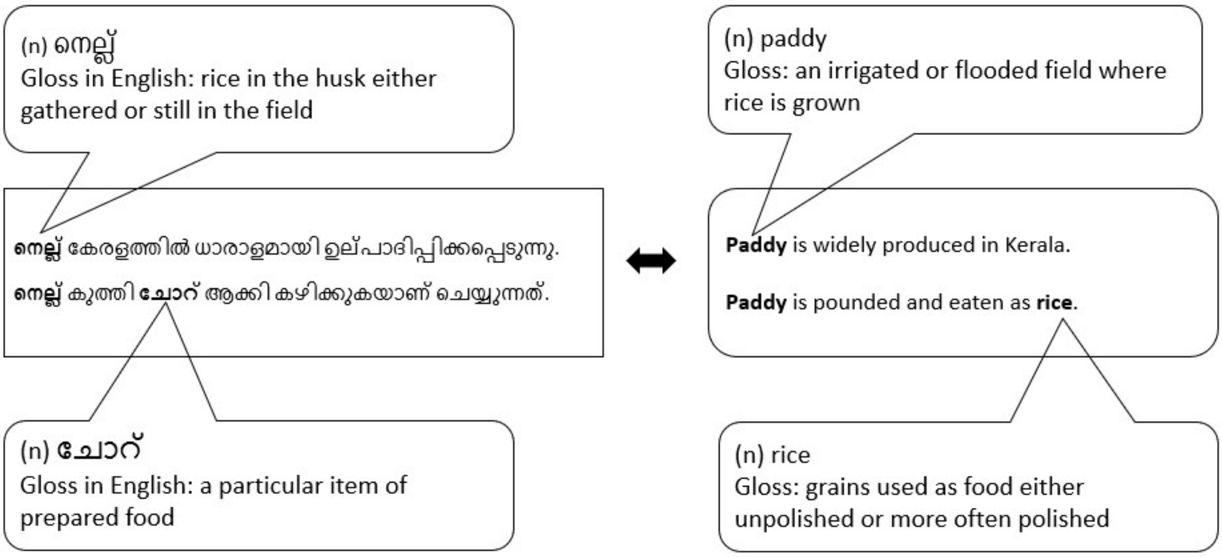

(n) നെല്ല്
Gloss in English: rice in the husk either gathered or still in the field
നെല്ല് കേരളത്തിൽ ധാരാളമായി ഉല്പാദിപ്പിക്കപ്പെടുന്നു.
നെല്ല് കുത്തി ചോറ് ആക്കി കഴിക്കുകയാണ് ചെയ്യുന്നത്.
(n) ചോറ്
Gloss in English: a particular item of prepared food
⬌
(n) paddy
Gloss: an irrigated or flooded field where rice is grown
Paddy is widely produced in Kerala.
Paddy is pounded and eaten as rice.
(n) rice
Gloss: grains used as food either unpolished or more often polished

Figure 6.3: Sentence translation from Malayalam to English

The concept from one language gets generalized due to the unavailability of that concept in a reference language, reducing the presence of minor culturally rich languages. Moreover, another issue is that translated sentence provides a different meaning. For example, in Figure 6.3, two sentences from Malayalam are translated into English. The first sentence is "നെല്ലു കേരളത്തി□ ധാരാളമായി ഉല്പാദിപ്പിക്കപ്‌പെടുന്നു (nellu keralathil dharalamayi ulpadikkapedunnu)," which means "grain of paddy is widely produced in Kerala." However, the online translation is that "paddy is widely produced in Kerala." So, the word "നെല്ലു (nellu)" is considered as paddy in the translation. The translation is confusing as "paddy" means "an irrigated or flooded field where rice is grown" from PWN. The word "rice" is not less used in languages, but the quality of translation is poor for under-resourced languages like Malayalam. We show that the translation system repeats the error in the second sentence. The second sentence is "നെല്ലു കുത്തി ചോറ് ആക്കിയാണ് കഴിക്കുകയാണ് ചെയ്യുന്നത് (nellu kuthi choru akki kazhikkukayanu cheyyunnat)," and means "grain of paddy is pounded and eaten as rice." The second sentence is translated as "paddy being pounded and eaten as rice." The example mentioned above shows the challenges of resources for under-resourced languages.

6.5 Summary

This chapter aims to show the problems Indian languages face without preservation and the diversity-aware. We show examples from the context of Indian languages that are culturally and linguistically rich. However, this could occur with most under-resourced languages. The examples used in this chapter from Amazon and Google Translate are all recent searches in the Malayalam language. The examples also point out the effect of the digital language divide on Malayalam.

Part III

Solution

Chapter 7

Proposed Methodology

7.1 Introduction

We have explained the importance of having a multilingual resource that helps preserve the languages by avoiding the challenges mentioned above. This chapter gives an overview of the proposed methodology, focusing on improving the existing approaches to building multilingual wordnet. There are two famous approaches to building wordnets: expansion and merging. The expansion approach uses reference wordnet and then translates the lexical elements into the target language. However, the expansion approach does not consider the target language concepts. In the merge, we have already developed a wordnet or some resource like a dictionary or thesaurus, and we will align the concepts with another fully developed reference wordnet. Our proposed approach is an extended merge approach. We use a large-scale multilingual independent resource as a reference for aligning the concepts and supporting the management of lexical gaps.

We check the validity of our proposed methodology for Indian languages; hence, we use the available and popular multilingual resource of Indian languages of IWN. We merge IWN with language-independent UKC' s universal lexical concepts. Our approach will develop a new structure to avoid the dominance of the language of Hindi. Furthermore, our methodology includes a stage for preserving the Indian languages by adding more concepts from diverse and culturally essential domains.

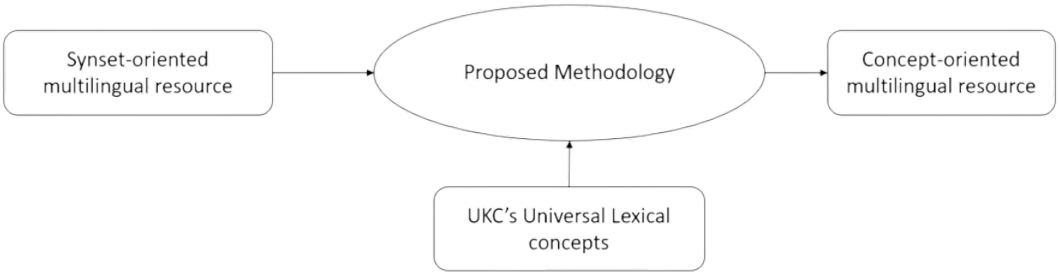

Synset-oriented multilingual resource
Proposed Methodology
Concept-oriented multilingual resource
UKC's Universal Lexical concepts

Figure 7.1: Overview of the proposed methodology

7.2 The Process

Figure 7.1 shows the overview of the proposed methodology. The methodology uses synset-oriented resources like IWN as input and produces a concept-oriented resource. The thesis considers the concepts-oriented resource as a resource that is not partial towards one language or culture. Our proposed approach employs UKC' s universal lexical concepts to make the resultant resource language independent. In addition, to re-use the concepts from input resources, our process will also consider how to add a new concept to the resultant resource. The methodology consists of four different phases, and they are in sequential order. A summary of the methodology phases is explained in the following sections.

7.2.1 Phase I - Design Resource

The first stage of our approach has three significant steps: quality check, filter and clean, and classify. This stage aims to understand the resource and evaluate how much effort someone needs to merge and make the resource concept-oriented. The term quality is subjective and ambiguous.

In this thesis, our definition of quality is checking if the resource is free from errors. Moreover, the thesis defines the errors that affect the cross-lingual mapping ability of the input resource. So, in the second step of this phase, we filter the part of the resource which needs linguistics knowledge and eventually clean the errors computationally that do not require human support. Then in the last step of the phase, we classify the rest of the resources into three groups based on the universal lexical concepts of the UKC. The groups we select are familiar, diverse, and language-specific concepts between the input resource and UKC' s lexical concepts. Chapter 8 gives more details about this phase.

7.2.2 Phase II - Common Concepts

The objective of this second phase of the proposed methodology is similar to the traditional merge approach; we merge the familiar concepts from the resources and validate the resultant resource using linguistics. We use a reference language from the input synset-oriented resource to select familiar concepts. The language selection will vary depending on the nature of the multilingual resource. We verify the mappings between the concepts from the input resources and the universal lexical concepts used by UKC. Based on the projected accuracy of equivalence mapping, we will adapt it to other languages to provide a concept-oriented multilingual resource. Chapter 9 provides more information on this stage.

7.2.3 Phase III - One-to-Many Concepts

The objective of the third stage of the methodology is to merge the one-to-many concepts. One-to-many concepts mean one concept from the source map to many concepts in the target language. That is a concept map in multilingual resources with a general or specific concept. Such concepts are available in the input synset-oriented resource but mapped using hypernymy or hyponymy relations. Hence, to merge the input resource, we need to find the correct equivalence mappings out of the set of mappings. We use language experts to find the correct equivalence. We then compare the responses from language experts to find the standard equivalence mappings. Furthermore, we apply the standard mappings to other languages and merge the concepts we linked to the universal lexical concepts added to the concept-oriented resource generated in the previous phase. Chapter 10 gives the details of this phase with examples

7.2.4 Phase IV - Language-specific Concepts

The objective of the fourth stage of the methodology is to merge the language-specific or culture-specific concepts. The thesis considers language-specific concepts as the synsets from a synset-oriented resource that do not have equivalence with universal lexical concepts. In this phase, consider the parents of synsets to merge the concepts in the final resource. We group the synsets according to the parent synset. Grouping of synsets helps to focus on a specific domain of concepts. So the linguists can translate the synsets into English from one group she has substantial knowledge. Translating the synsets into English helps to understand whether or not the concept already exists in the universal lexical concept. If it does not exist, the concept will indicate a lexical gap. Chapter 11 provides the details of the process.

Chapter 8

Phase I: Resource Design

8.1 Introduction

The proposed methodology aims to develop resources for preserving languages in the form of concept-oriented resources. The methodology's outcome will not depend on any languages or cultures. The methodology uses Indian languages to show the implementation, but we hope to test the methodology for other languages, especially those under-resourced. We consider a language as under-resourced if the language lacks resources or resources are in limited number, lacks language processing tools, and lacks language experts. We can inherit knowledge from different cultures by supporting the development of under-resourced languages.

Figure 8.1 shows the steps in phase I. Our approach starts with an assumption that the languages will at least have one existing lexical resource. We have not included the stage of developing, converting, or developing a lexical resource. Moreover, we are using language experts to ensure the quality of the outcome. Phase 1 stage is the preparation stage of the methodology to understand the input and make the process of merging the resource easier.

8.2 The Process

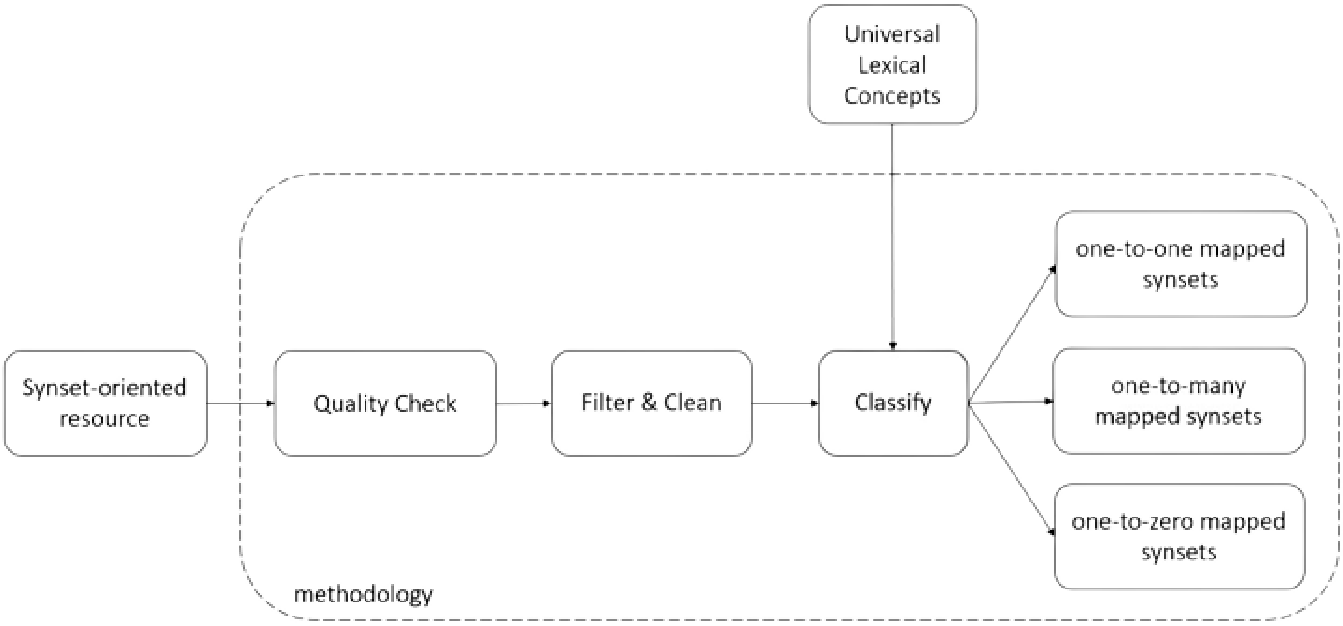


Figure 8.1: Phase I: Resource Design

The three steps in our approach are,

- **Quality check**: We check the quality of the input resource. We check if the resource has any errors. We aim to avoid poor quality resultant resources.
- **Filter and clean**: Next step is to filter the synsets with errors. Furthermore, we will fix some of the errors if possible since some of the errors we find will fix themselves when we evolve the resultant resource.
- **Classify**: Once we filter and clean the resource, we classify the synsets to handle the merge approach of the resource with ease and straightforwardness.

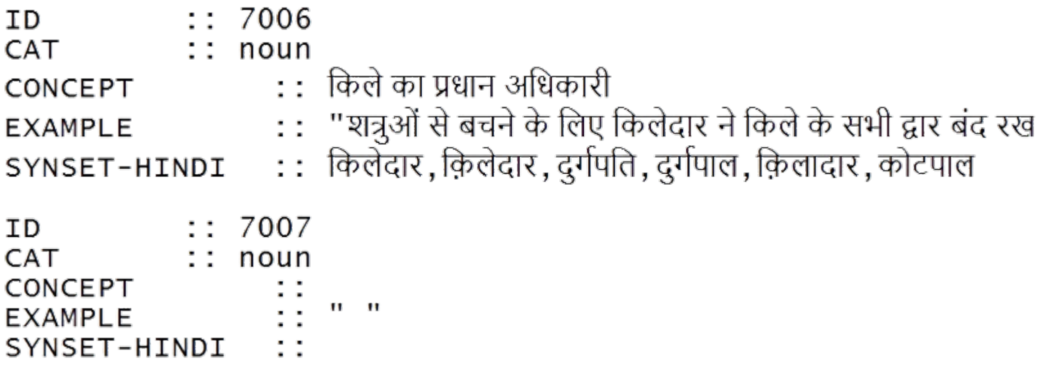
8.2.1 Quality Check

Quality assessment [19], quality assurance [44], and quality control [20] are three terms used interchangeably in the translation community to refer to quality-related initiatives.

- Measuring a product's compatibility with quality criteria is known as **quality assessment** or **quality evaluation** [27].
- **Quality assurance** refers to methods for preventing item errors or flaws and avoiding issues when providing solutions. Continuous quality assessment is essential for quality assurance [27].
- The practice of checking whether manufactured products fulfill stated quality criteria is known as **quality control** [27].

We estimated the lexical resource's quality based on the study [31]. We began the process by splitting the data into a sample set to understand it better. We reviewed the IWN to understand the resource metadata better and look for any mistakes. We have broken down the errors we found into categories below,

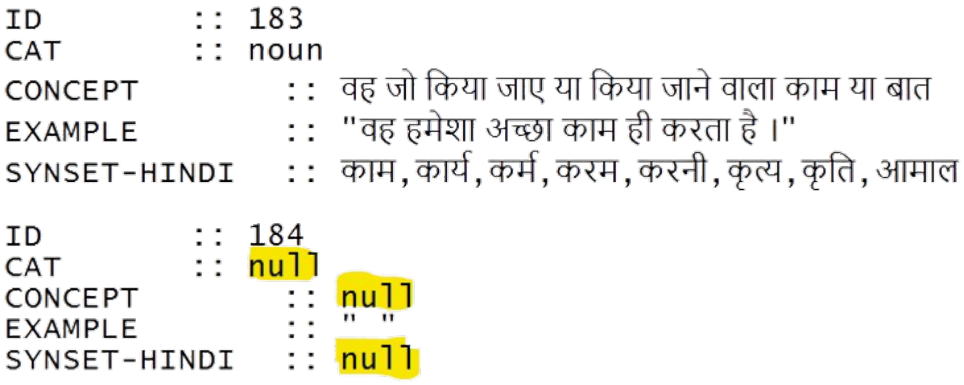
- Type 1 - Empty example, gloss, and synset: the fields example, gloss, and synset have no value. Data is attached to the other fields causing this error. Figure 8.2 shows an example of the Type 1 error. The fields for the synset **ID 7007** are empty in the Hindi WordNet of IWN. Except for the part of speech that indicated using the label **CAT**, all the fields for the synset are empty. The error may indicate that the synset for ID 7007 exists for other wordnets in IWN but not for Hindi. We discovered that Type 1 errors are too many in Hindi, which we assume is one of the methods utilized to connect all wordnets.



```
ID          :: 7006
CAT         :: noun
CONCEPT            :: किले का प्रधान अधिकारी
EXAMPLE            :: "शत्रुओं से बचने के लिए किलेदार ने किले के सभी द्वार बंद रख
SYNSET-HINDI       :: किलेदार,क़िलेदार,दुर्गपति,दुर्गपाल,क़िलादार,कोटपाल

ID          :: 7007
CAT         :: noun
CONCEPT            ::
EXAMPLE            :: " "
SYNSET-HINDI       ::
```

Figure 8.2: Example for Type 1

- Type 2 - Null example, gloss, and synset: the fields example, gloss, and synset have no value but the label “null.” Figure 8.3 shows an example of the Type 2 error in the Hindi WordNet of IWN. For the synset ID 184, the fields, part of speech, gloss, and synset have the value null, and the field example has an empty value. This error is similar to Type 1, except the field has a label, but we observed this error in Hindi.

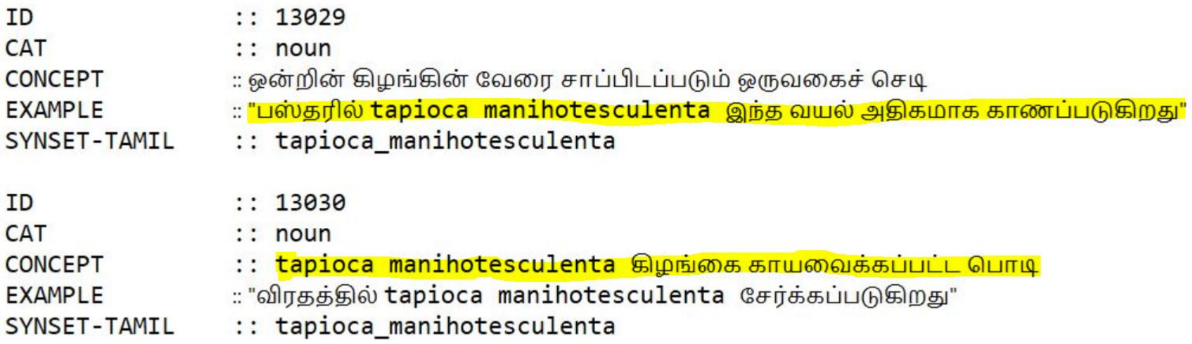


```
ID          :: 183
CAT         :: noun
CONCEPT            :: वह जो किया जाए या किया जाने वाला काम या बात
EXAMPLE            :: "वह हमेशा अच्छा काम ही करता है ।"
SYNSET-HINDI       :: काम,कार्य,कर्म,करम,करनी,कृत्य,कृति,आमाल

ID          :: 184
CAT         :: null
CONCEPT            :: null
EXAMPLE            :: " "
SYNSET-HINDI       :: null
```

Figure 8.3: Example for Type 2

- Type 3 - Random characters in the fields: some characters are part of the value. This error gives incorrect data if we do not remove it. Figure 8.4 shows an example of the Type 3 error. For the synsets with

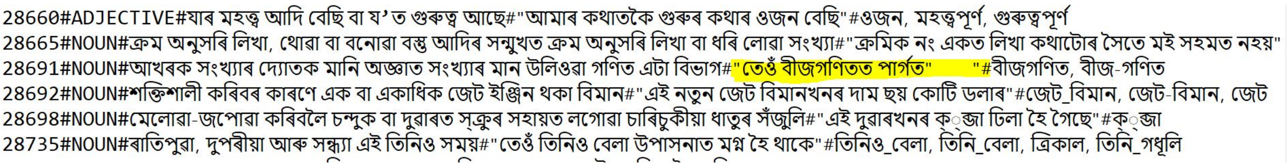
IDs 13029 and 13030 in Tamil wordnet, gloss and synset fields have random characters on the part of the data.



```
ID              :: 13029
CAT             :: noun
CONCEPT         :: ஒன்றின் கிழங்கின் வேரை சாப்பிடப்படும் ஒருவகைச் செடி
EXAMPLE         :: "பஸ்தரில் tapioca manihotesculenta இந்த வயல் அதிகமாக காணப்படுகிறது"
SYNSET-TAMIL    :: tapioca_manihotesculenta

ID              :: 13030
CAT             :: noun
CONCEPT         :: tapioca manihotesculenta கிழங்கை காயவைக்கப்பட்ட பொடி
EXAMPLE         :: "விரதத்தில் tapioca manihotesculenta சேர்க்கப்படுகிறது"
SYNSET-TAMIL    :: tapioca_manihotesculenta
```

Figure 8.4: Example for Type 3 in the gloss field

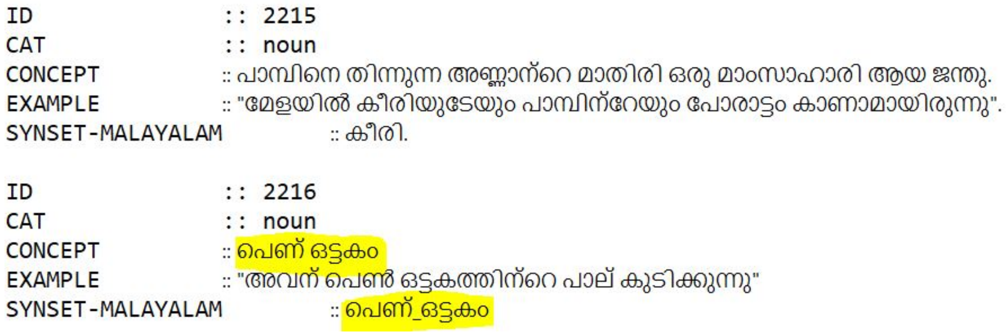
- Type 4 - Presence of extra double quotes: The field example shows within double quotes (" "). Type 4 errors occur when an additional double quote is present in any of the fields of a synset. So this type of error is essential to address because it will misinterpret the system; that is, additional double quotes give the idea of an additional field. Figure 8.5 shows an example of the Type 4 error. Figure 8.5 highlights the additional double quotes for the synset ID 28691.



```
28660#ADJECTIVE#যাৰ মহত্ত্ব আদি বেছি বা য’ত গুৰুত্ব আছে#"আমাৰ কথাতকৈ গুৰুৰ কথাৰ ওজন বেছি"#ওজন, মহত্ত্বপূৰ্ণ, গুৰুত্বপূৰ্ণ
28665#NOUN#ক্ৰম অনুসৰি লিখা, থোৱা বা বনোৱা বস্তু আদিৰ সন্মুখত ক্ৰম অনুসৰি লিখা বা ধৰি লোৱা সংখ্যা#"ক্ৰমিক নং একত লিখা কথাটোৰ সৈতে মই সহমত নহয়"
28691#NOUN#আখৰক সংখ্যাৰ দ্যোতক মানি অজ্ঞাত সংখ্যাৰ মান উলিওৱা গণিত এটা বিভাগ#"তেওঁ বীজগণিতত পাৰ্গত"       "#বীজগণিত, বীজ-গণিত
28692#NOUN#শক্তিশালী কৰিবৰ কাৰণে এক বা একাধিক জেট ইঞ্জিন থকা বিমান#"এই নতুন জেট বিমানখনৰ দাম ছয় কোটি ডলাৰ"#জেট_বিমান, জেট-বিমান, জেট
28698#NOUN#মেলোৱা-জপোৱা কৰিবলৈ চন্দুক বা দুৱাৰত স্ক্ৰুৰ সহায়ত লগোৱা চাৰিচুকীয়া ধাতুৰ সঁজুলি#"এই দুৱাৰখনৰ ক্‌ ্জা ঢিলা হৈ গৈছে"#ক্‌ ্জা
28735#NOUN#ৰাতিপুৱা, দুপৰীয়া আৰু সন্ধ্যা এই তিনিও সময়#"তেওঁ তিনিও বেলা উপাসনাত মগ্ন হৈ থাকে"#তিনিও_বেলা, তিনি_বেলা, ত্ৰিকাল, তিনি_গধূলি
```

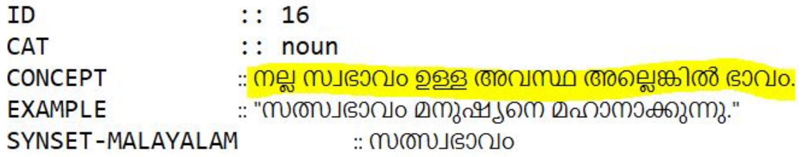
Figure 8.5: Example for Type 4

- Type 5 –Gloss is the same as synsets: the fields gloss and synset values are the same. Figure 8.6 shows an example of a Type 5 error in the Malayalam language. Figure 8.6 shows the field gloss presented with the label CONCEPT and synset with SYNSET-MALAYALAM. We highlight both identical values.



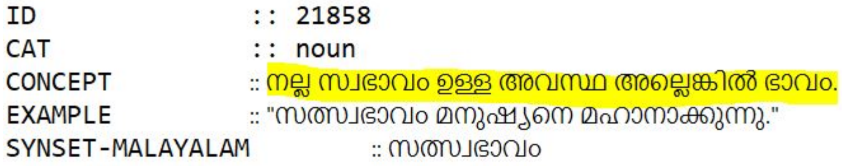
```
ID                  :: 2215
CAT                 :: noun
CONCEPT             :: പാമ്പിനെ തിന്നുന്ന അണ്ണാന്റെ മാതിരി ഒരു മാംസാഹാരി ആയ ജന്തു.
EXAMPLE             :: "മേളയിൽ കീരിയുടേയും പാമ്പിന്റേയും പോരാട്ടം കാണാമായിരുന്നു".
SYNSET-MALAYALAM          :: കീരി.

ID                  :: 2216
CAT                 :: noun
CONCEPT             :: പെണ് ഒട്ടകം
EXAMPLE             :: "അവന് പെൺ ഒട്ടകത്തിന്റെ പാല് കുടിക്കുന്നു"
SYNSET-MALAYALAM          :: പെണ്_ഒട്ടകം
```

Figure 8.6: Example for Type 5



```
ID                  :: 16
CAT                 :: noun
CONCEPT             :: നല്ല സ്വഭാവം ഉള്ള അവസ്ഥ അല്ലെങ്കിൽ ഭാവം.
EXAMPLE             :: "സത്സ്വഭാവം മനുഷ്യനെ മഹാനാക്കുന്നു."
SYNSET-MALAYALAM          :: സത്സ്വഭാവം
```

(a) For synset id 16



```
ID                  :: 21858
CAT                 :: noun
CONCEPT             :: നല്ല സ്വഭാവം ഉള്ള അവസ്ഥ അല്ലെങ്കിൽ ഭാവം.
EXAMPLE             :: "സത്സ്വഭാവം മനുഷ്യനെ മഹാനാക്കുന്നു."
SYNSET-MALAYALAM          :: സത്സ്വഭാവം
```

(b) For synset id 21858

Figure 8.7: Example for duplicate gloss

Duplicate gloss error is another type of error we analyzed. We need to estimate this inaccuracy for other languages in the future because it can only estimate with the help of a native speaker. The example of a duplicate gloss error is seen in Figure 8.7, where synset IDs 16 and 21858 have the same gloss. The error led to various conclusions about the resource's quality and the language's diversity. The IWN team could not discover a correct synonym value, so they offered the concept's definition or could not find the synset's definition. The gloss is replaced with the synset value until they can find one in the future, and the resource will not appear incomplete.

8.2.2 Filter and Clean

In IWN, we list all languages' errors defined in the previous step. The details of the errors can be found in the chapter's results section. Type 1 and Type 2 errors are instances of adding concepts, so one of the future tasks of evolving resources. After assessing each language's errors, we filter the entries from IWN with Type 1, Type 2, and Type 5 errors. Type 3 and Type 4 can be corrected manually. Once we clean the records with Type 3 and Type 4, we will consider the filtered and cleaned resource as our synset-oriented resource for our next step. Figure 8.8 shows the filter and clean stage summary.

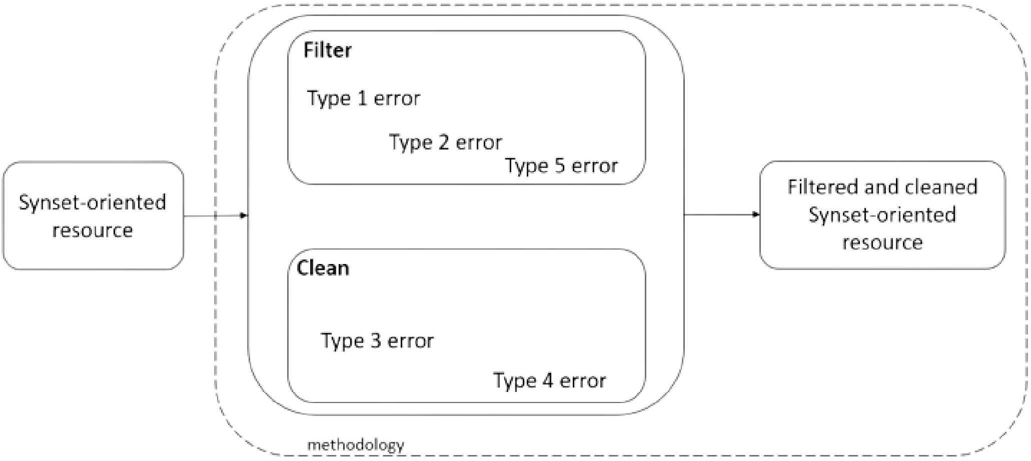


Figure 8.8: Filter and clean stage

8.2.3 Merging

The proposed approach differs from the standard merge method for developing wordnets and uses an additional resource, UKC, as explained in the work [32]. The UKC features a concept layer fully separated from the language, which utilizes PWN synsets as concepts for the PWN hierarchy [15]. As a result, the IWN synset aligned with the UKC concepts will be associated with the matching PWN synset. Also, it helps the UKC generate new UKC ids for those IWN synsets which do not correspond to UKC (and therefore) to PWN. In the UKC, concepts are associated with unique ids and connected to language in three possible ways [16].

1. The concept id is mapped to a synset id (one-to-one), indicating that the concept is lexicalized in that language.
2. The concept id has been designated as a lexical gap in that language, indicating that it has not been lexicalized.
3. The concept id is not linked to anything.

The merging stage used the error-free resource generated during the analysis phase as input. We divided synsets in the merged resource into three groups using the concept layer principle.

- Group A: One synset from IWN has a corresponding single concept in UKC. These are the IWN synsets that have a one-to-one mapping with PWN. Figure 8.9 shows the example of a Group A synset. In Figure 8.9, the English synset “deeds” has gloss “performance of moral or religious acts” has one corresponding synset in Hindi, “सत्कर्म ” (*sathkarm*), with gloss “ऐसा कार्य जो नीतिपरक हो ” (*isa kary jo neethipark ho*) and has one corresponding synset in Malayalam, “സത്കര്മ്മം ” (*sathkarmam*) with gloss “നീതി യുക്തമായ കര്മ്മം ” (*neethi yukthamaya karmam*). Figure 8.9 explains one-to-one mappings between English, Hindi, and Malayalam.
- Group B: Many synsets from IWN have a corresponding single concept in the UKC. These are the IWN synsets that map to PWN in a one-to-many relationship. Figure 8.10 shows an example of Group B where the English synset “achievement” has gloss “the action of accomplishing something” and has two corresponding synsets in Hindi

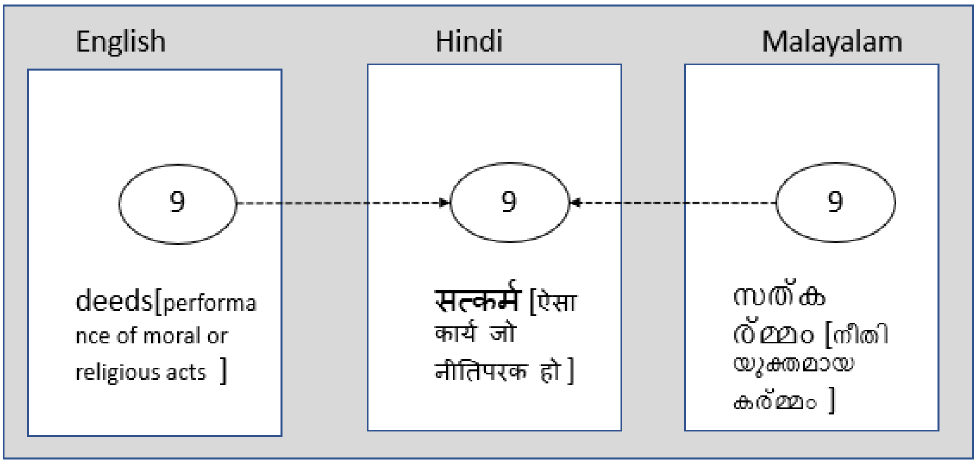

English
Hindi
Malayalam
9
9
9
deeds[performance of moral or religious acts]
सत्कर्म [ऐसा कार्य जो नीतिपरक हो]
സത്കർമ്മം [നീതിയുക്തമായ കർമ്മം]

Figure 8.9: Example for Group A synset

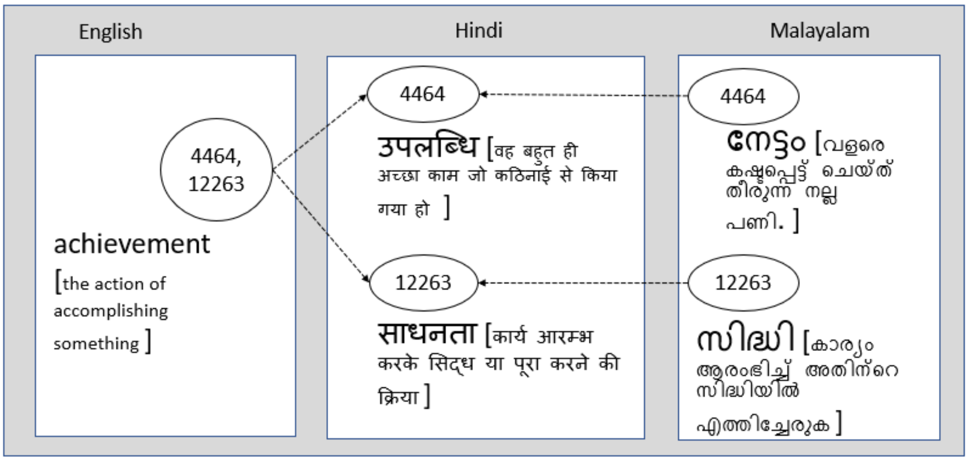

English
Hindi
Malayalam
4464, 12263
achievement [the action of accomplishing something]
4464
उपलब्धि [वह बहुत ही अच्छा काम जो कठिनाई से किया गया हो]
12263
साधनता [कार्य आरम्भ करके सिद्ध या पूरा करने की क्रिया]
4464
നേട്ടം [വളരെ കഷ്ടപ്പെട്ട് ചെയ്ത് തീരുന്ന നല്ല പണി.]
12263
സിദ്ധി [കാര്യം ആരംഭിച്ച് അതിന്റെ സിദ്ധിയിൽ എത്തിച്ചേരുക]

Figure 8.10: Example for Group B synsets

and Malayalam. The first one with ID 4464 is mapped to Hindi synset “उपलब्धि” (*upalabdhi*) with gloss “वह बहुत ही अच्छा काम जो कठिनाई से किया गया हो” (*vah bahuth hi acha kam jo katinay se kiya gaua ho*) and to Malayalam synset “നേട്ടം” (*nettam*) with gloss “വളരെ കഷ്ടപ്പെട്ട് ചെയ്ത് തീരുന്ന നല്ല പണി” (*valare kashtapettu cheyt theerunna pani*). The second one with ID 12263 is mapped to Hindi synset “साधनता” (*sathantha*) with gloss “कार्य आरम्भ करके सिद्ध या पूरा करने की क्रिया” (*kary arambh karke sidh ya poora karne ke kriya*) and to Malayalam synset “സിദ്ധി” (*sidhi*) with gloss “കാര്യം ആരംഭിച്ച് അതിന്റെ സിദ്ധിയിൽ

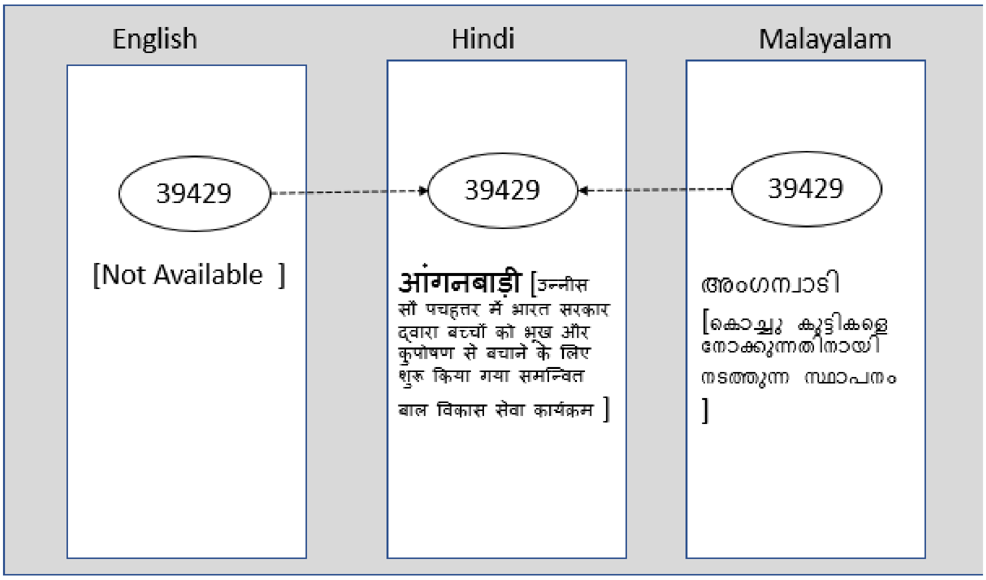

English
Hindi
Malayalam
39429
39429
39429
[Not Available]
आंगनबाड़ी [उन्नीस सौ पचहत्तर में भारत सरकार द्वारा बच्चों को भूख और कुपोषण से बचाने के लिए शुरू किया गया समन्वित बाल विकास सेवा कार्यक्रम]
അംഗന്വാടി
[കൊച്ചു കുട്ടികളെ നോക്കുന്നതിനായി നടത്തുന്ന സ്ഥാപനം]

Figure 8.11: Example for Group C synset

എത്തിച്ചേരുക" (*karyam arambhichu athinte sidhiyil ethi cheruka*). The presence of hypernymy links [23] cause such language mappings to diminish the resource' s ability to be extended.

- Group C : In UKC, there is no concept for such an IWN synset. These are the IWN synsets that have one to zero mapping with PWN. Figure 8.11, shows an example of Group C. A Hindi synset "आंगनबाड़ी" (*anganbadi*) with gloss "उन्नीस सौ पचहत्तर में भारत सरकार द्वारा बच्चों को भूख और कुपोषण से बचाने के लिए शुरू किया गया समन्वित बाल विकास सेवा कार्यक्रम " (*unnis sou pachhathar mem bharat sarkar dhwara bacchom kho fhookh aur krposhna se bachane ke liye shuru kiya gaya samnivth vikas seva karykremu*) has corresponding synset in Malayalam, "അംഗന്വാടി" (*anganvadi*) with gloss "കൊച്ചു കുട്ടികളെ നോക്കുന്നതിനായി നടത്തുന്ന സ്ഥാപനം" (*kochu kuttikale nokkunnathanayi nadathunna sthapanam*). However, it does not have a corresponding synset for English wordnet in IWN. The concept mentioned above is Indian terminology for a location to look after children; it differs from "daycare," meaning "childcare during the day while parents work." The Indian concept emerged into the language due to an Indian government's mission to promote children's welfare. The absence of lexicalization is better since it could have mismapped to a general concept like "shelter," which would lead to the loss of the concept in the future.

8.3 Results

In Table 8.1, we calculated the mistakes for each language. As we can see, Type 1 is the most common mistake type in languages. If Type 1 appears in any record fields, it is considered. We carefully reviewed several languages with over 10K Type 1 errors and found that they had empty example fields. The gloss or synset fields are used to add the example field. As a result, the Oriya wordnet has the most mistakes. As a result, we had to correct them manually. However, we do not recommend using the method on any of the resource's other wordnets. In the IWN, the Telugu wordnet has fewer mistakes. From Table 8.1, we can observe that the Malayalam WordNet required much attention. Type 2 and Type 5 errors are higher in Malayalam.

The presence of random characters is higher in the Sanskrit wordnet. Hindi WordNet has the highest presence of double quotes.

Table 8.2 lists the number of synsets in each group based on the UKC' s universal lexical elements. Hindi WordNet has the highest number of synsets, and Nepali wordnet has fewer. Group A and Group B synsets represent the corresponding lexicalized concepts. Group A shows the common concepts between the Indian language and English. As we can see, there are variations in each language. For example, Hindi has the most number of synsets. However, only 28 percent of the synsets are familiar, but for Nepali, 53 percent of the synsets are familiar. Group B represents

Table 8.1: Mistakes identified in IndoWordNet

Sl No	Languages	Total synsets	Type 1	Type 2	Type 3	Type 4	Type 5	Total errors
1	Assamese	14958	305	0	1	0	0	306
2	Bengali	36346	16491	0	0	0	7	16498
3	Bodo	15785	3	10	0	0	11	24
4	Gujarati	35599	12642	58	0	0	2	12702
5	Hindi	40371	581	38	0	393	0	1012
6	Kannada	22042	14	114	0	0	1	129
7	Kashmiri	29469	14279	27	0	0	6	14312
8	Konkani	32370	7585	107	0	0	3	7695
9	Malayalam	30140	1970	2778	0	0	2374	5152
10	Manipuri	16351	14	29	0	0	1	44
11	Marathi	32721	14	0	2	0	0	16
12	Nepali	11713	2169	0	0	0	5	2174
13	Oriya	35284	35266	0	1	0	0	35267
14	Punjabi	32364	18551	0	1	0	11	18563
15	Sanskrit	38070	595	0	15	0	0	610
16	Tamil	25419	62	0	7	0	521	590
17	Telugu	21091	0	0	1	0	2	3
18	Urdu	34280	11412	0	2	0	2	11416

the mappings of diverse concepts. For Hindi, 33 percent of the synsets are group B, and only 12 percent of Nepali synsets are group B. Group C indicates the language-specific concepts or concepts that need to be mapped. For Hindi, around 39 percent of the synsets are group C, and around 35 percent of Nepali synsets are group C. We can see a pattern for group C synsets.

8.4 Summary

The chapter is dedicated to designing the input resource for transforming the resource structure using the UKC' s lexical concepts and creating a concept-oriented resource. The objective of this chapter is to prepare the

Table 8.2: Statistics of synsets in three groups for Indian languages

No	**Languages**	**#IWN synsets**	**#Group A**	**#Group B**	**#Group C**
1	Assamese	14958	7314	2646	4998
2	Bengali	36346	9428	10020	16898
3	Bodo	15785	7346	2838	5601
4	Gujarati	35599	9342	9899	16358
5	Hindi	40371	11283	13416	15672
6	Kannada	22042	7880	4661	9501
7	Kashmiri	29469	8544	8360	12565
8	Konkani	32370	9218	9413	13739
9	Malayalam	30140	9719	10127	10294
10	Manipuri	16351	7328	2855	6168
11	Marathi	32721	8691	11265	12765
12	Nepali	11713	6218	1505	3990
13	Oriya	35284	9281	9844	16159
14	Punjabi	32364	9123	8548	14693
15	Sanskrit	38070	7213	22481	8376
16	Tamil	25419	9745	7123	8551
17	Telugu	21091	7633	4303	9155
18	Urdu	34280	9205	8949	16126

resource to merge. Furthermore, we ensured the quality of the resultant resource by identifying the errors and fixing them. Still, the languages have more than 10K synsets, so the manual process takes more time. Hence, to simplify the process, we grouped the synsets using the UKC’ s principle, and we will process the resource by taking each group of synsets defined in this chapter in the coming chapters.

Chapter 9

Phase II : Merge Common Concepts

9.1 Introduction

This chapter aims to demonstrate how to extract synsets from a cleaned input resource and translate them into concepts in a concept-oriented resource; this is how we integrate previously mapped concepts. If the concepts have previously been mapped, the thesis assumes that lexicalization is possible from one language to another. Furthermore, lexicalization is only possible for concepts shared between the source and target languages if the concept from the source language is also available in the target language. The chapter explains a straightforward step for merging a multilingual resource.

We prepared the three synset modules for transfer into the new resource structure in the previous chapter. In this chapter, we will look at group A, or one-to-one mapped synsets, the first module of synsets.

9.2 The Process

The process of merging common concepts consists of five steps. We first select a lexical reference resource from the cleaned and grouped input synset-oriented multilingual lexical resource. Language selection is easier for a

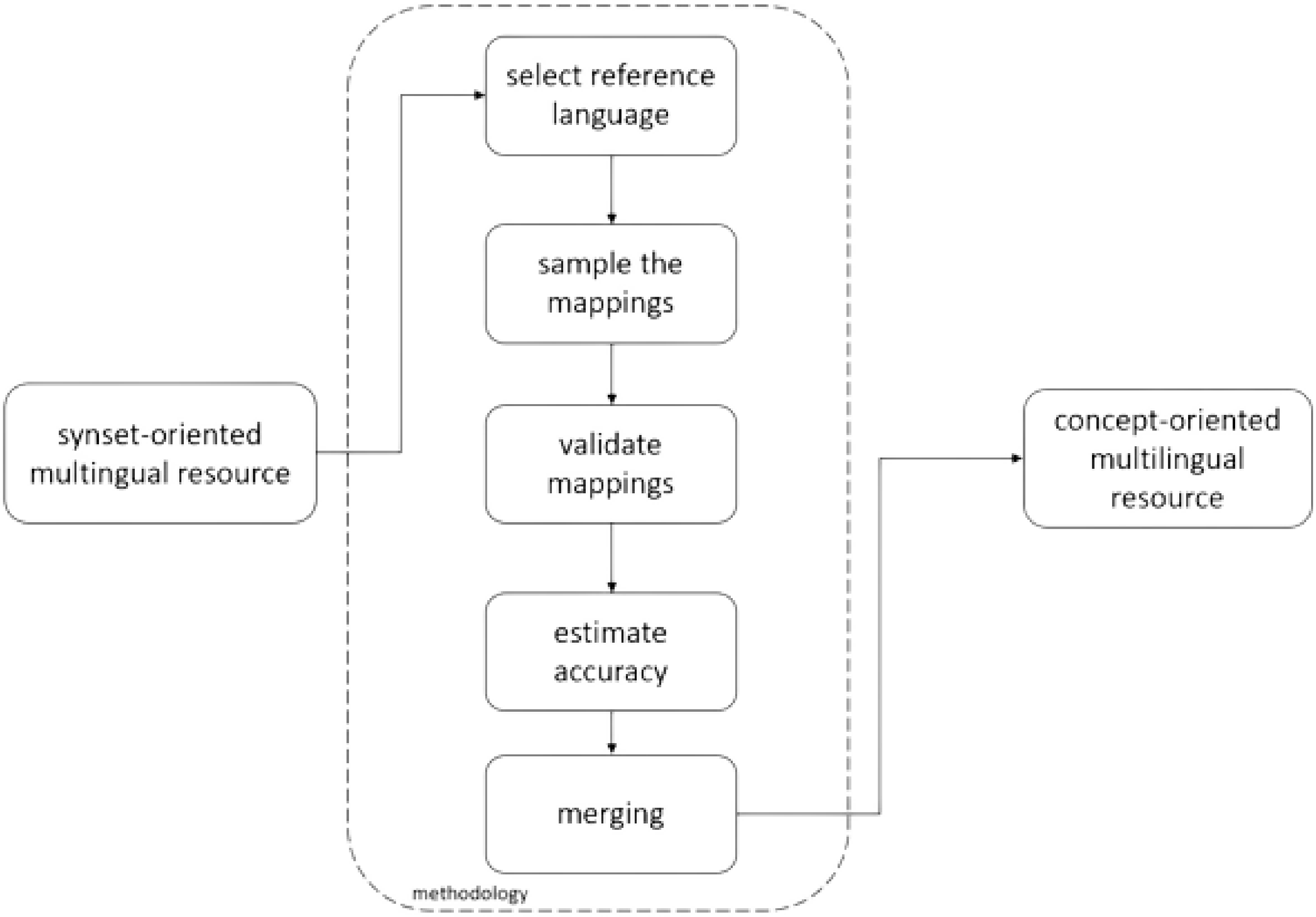

select reference language
sample the mappings
validate mappings
estimate accuracy
merging
synset-oriented multingual resource
concept-oriented multilingual resource
methodology

Figure 9.1: Phase II: Merge Common Concepts

multilingual resource since it uses a reference language to connect the resource. We use the common concepts between the selected language and the UKC' s lexical concepts, which we grouped in the previous chapter. Once we select the reference language, we will sample the whole group A synsets. We validate the sample of one-to-one mapped synsets with the help of experts in the selected language. We estimate the accuracy of the equivalence mappings in group A synsets and, based on the accuracy, merge the synsets from the input resource. Figure 9.1 shows the steps in the process.

9.2.1 Selection of Language

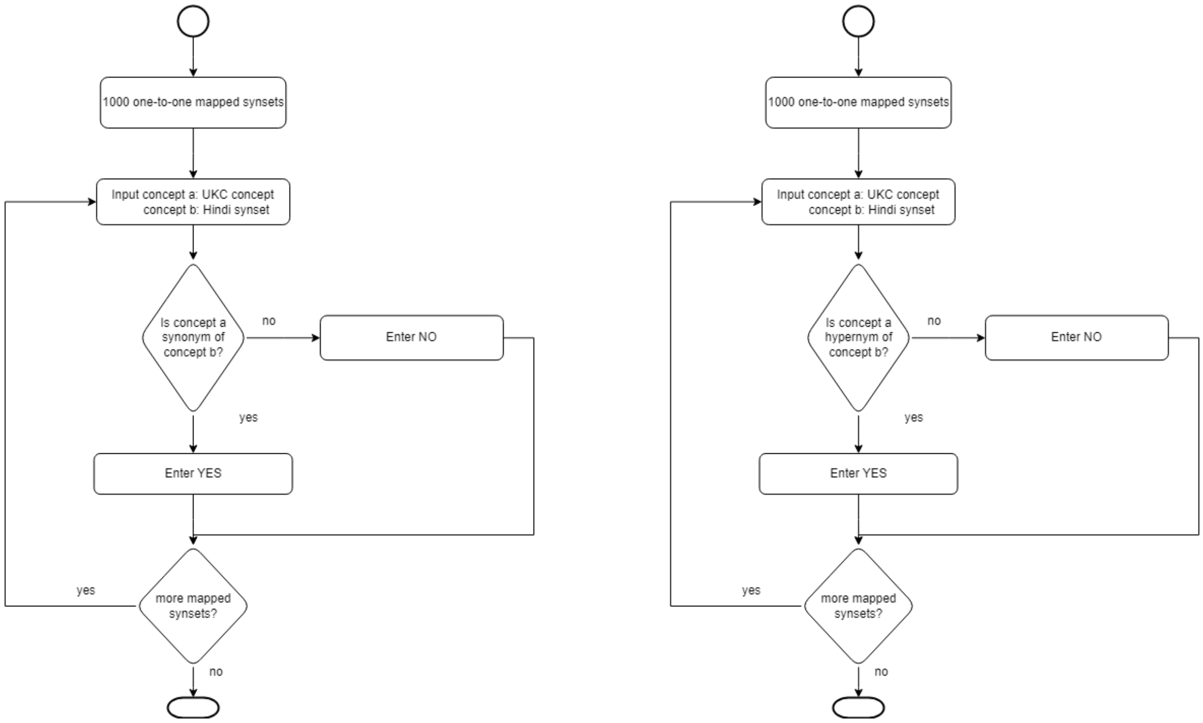
The objective of the language selection step is to make resource merging easier. The reference language used in a multilingual lexical resource like EWN and OMW are English, and for IWN, Hindi is the reference language. Before we decided to take Hindi, we experimented with another language, Malayalam. However, the IWN development is based on the Hindi language. The IWN team collaborated with universities across India and linked other language wordnets with Hindi. Hence, taking Hindi as a reference language is an ideal choice.

Also, selecting Hindi as a reference language helps us include more concepts in our resultant resource since Hindi has more synsets than other languages in IWN. This chapter focuses on merging the common concepts between English and Hindi. Moreover, Hindi is from the same linguistic family as English, which could be an advantage for the resultant resource.

9.2.2 Sampling of the Resource

The goal is to merge the one-to-one mapped synsets. In our case, we are extracting the one-to-one mapped synsets from an existing resource. Hence we need only to validate the mappings. We selected one reference language in the previous section and the selected language has more than 10K one-to-one mapped synsets. Ideally, we would like to analyze every part of the mapped synsets to measure the correctly mapped ones accurately; however, the process is both time-consuming, expensive, and labor-intensive, so it is not economically feasible to analyze large amounts of resources. Therefore, it is a standard practice to select a fraction of the whole material for analysis and assume that its properties represent the complete resource.

Instead of checking the equivalent mappings of the entire Hindi lan-


1000 one-to-one mapped synsets
Input concept a: UKC concept
concept b: Hindi synset
Is concept a synonym of concept b?
no
Enter NO
yes
Enter YES
more mapped synsets?
yes
no
1000 one-to-one mapped synsets
Input concept a: UKC concept
concept b: Hindi synset
Is concept a hypernym of concept b?
no
Enter NO
yes
Enter YES
more mapped synsets?
yes
no

(a) Synonymy validation

(b) Hypernymy validation

Figure 9.2: Flow diagram for validation tasks between UKC and Hindi

guage, we have decided to sample the synsets. There are around 40371 synsets available in Hindi WordNet. Hindi has 11283 one-to-one mapped synsets, and we prepared a sample of 1000 one-to-one mapped synsets.

9.2.3 Validating the Mapping

This step gives the language experts the sampled synsets of equivalence mappings. Figure 9.4, Figure 9.5, and Figure 9.6 show the snapshot of the file we have given to collect the experts' responses. We validate our mappings with a native speaker who understands English or a linguistic expert who can give us comments based on etymology. The expert must validate two types of relations in this mapping by answering the three questions: similarity (synonym) and parent-child (hypernym). The first two questions consider concept **"a"** as a UKC concept and concept **"b"** as a Hindi synset.

- Is the concept$a$ a synonym of concept b?
- Is the concept a a hypernym of concept b?

The third question considers concept a and concept b from Hindi, and the validator has to answer the question,

- Is the concept$a$ a hypernym of concept b?

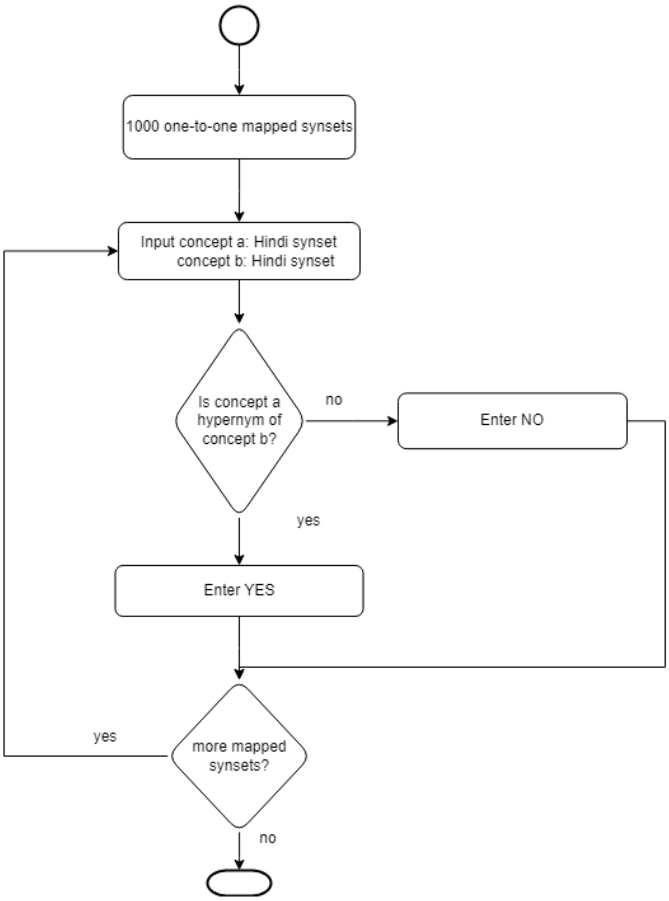
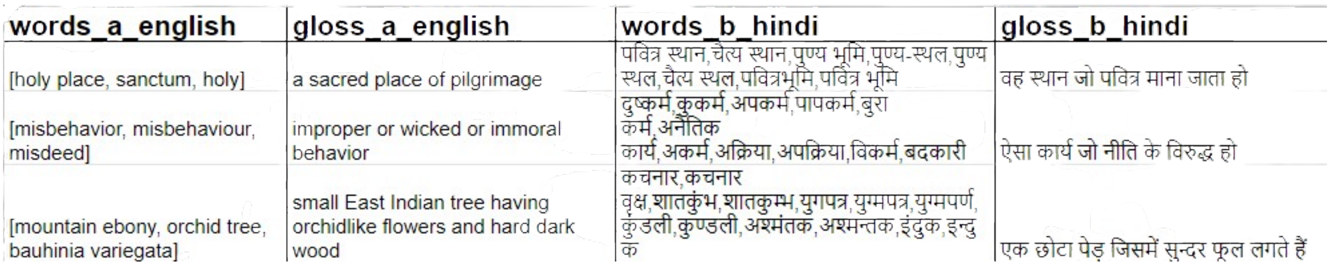

1000 one-to-one mapped synsets
Input concept a: Hindi synset
concept b: Hindi synset
Is concept a hypernym of concept b?
no
Enter NO
yes
Enter YES
yes
more mapped synsets?
no

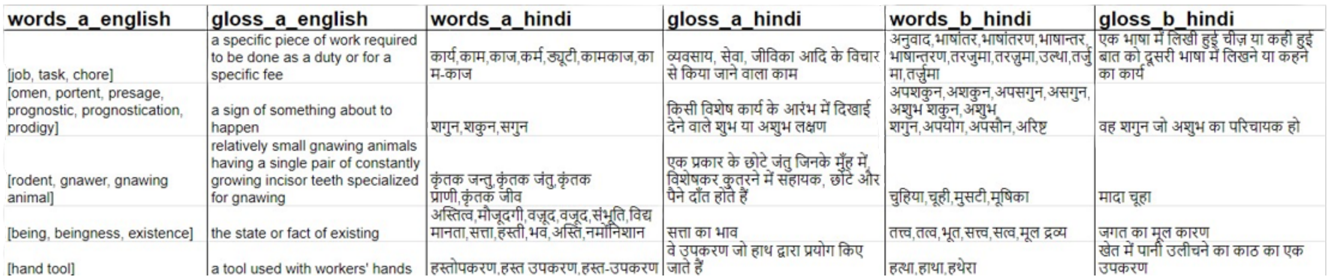
Figure 9.3: Flow diagram for validation task within Hindi: hypernymy validation

Validation has to respond to the questions mentioned above in three Excel files. Figure 9.4 shows the snapshot of the excel file for synonymy validation between the UKC and reference language. To validate the hypernymy, we use the excel file shown in Figure 9.5 between the UKC and reference language and Figure 9.6 for validating hypernymy within the reference language.



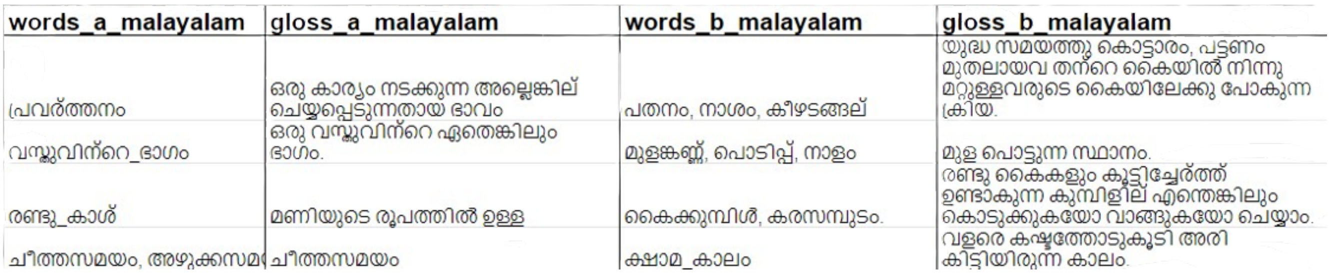
words_a_english	gloss_a_english	words_b_hindi	gloss_b_hindi
[holy place, sanctum, holy]	a sacred place of pilgrimage	पवित्र स्थान,चैत्य स्थान,पुण्य भूमि,पुण्य-स्थल,पुण्य स्थल,चैत्य स्थल,पवित्रभूमि,पवित्र भूमि	वह स्थान जो पवित्र माना जाता हो
[misbehavior, misbehaviour, misdeed]	improper or wicked or immoral behavior	दुष्कर्म,कुकर्म,अपकर्म,पापकर्म,बुरा कर्म,अनैतिक कार्य,अकर्म,अक्रिया,अपक्रिया,विकर्म,बदकारी	ऐसा कार्य जो नीति के विरुद्ध हो
[mountain ebony, orchid tree, bauhinia variegata]	small East Indian tree having orchidlike flowers and hard dark wood	कचनार,कचनार वृक्ष,शातकुंभ,शातकुम्भ,युगपत्र,युग्मपत्र,युग्मपर्ण,कुंडली,कुण्डली,अश्मंतक,अश्मन्तक,इंदुक,इन्दुक	एक छोटा पेड़ जिसमें सुन्दर फूल लगते हैं

Figure 9.4: Excel file for synonymy validation between UKC and Hindi



words_a_english	gloss_a_english	words_a_hindi	gloss_a_hindi	words_b_hindi	gloss_b_hindi
[job, task, chore]	a specific piece of work required to be done as a duty or for a specific fee	कार्य,काम,काज,कर्म,ड्यूटी,कामकाज,काम-काज	व्यवसाय, सेवा, जीविका आदि के विचार से किया जाने वाला काम	अनुवाद,भाषांतर,भाषांतरण,भाषान्तर,भाषान्तरण,तरजुमा,तरज़ुमा,उल्था,तर्जुमा,तर्ज़ुमा	एक भाषा में लिखी हुई चीज़ या कही हुई बात को दूसरी भाषा में लिखने या कहने का कार्य
[omen, portent, presage, prognostic, prognostication, prodigy]	a sign of something about to happen	शगुन,शकुन,सगुन	किसी विशेष कार्य के आरंभ में दिखाई देने वाले शुभ या अशुभ लक्षण	अपशकुन,अशकुन,अपसगुन,असगुन,अशुभ शकुन,अशुभ शगुन,अपयोग,अपसौन,अरिष्ट	वह शगुन जो अशुभ का परिचायक हो
[rodent, gnawer, gnawing animal]	relatively small gnawing animals having a single pair of constantly growing incisor teeth specialized for gnawing	कृंतक जन्तु,कृंतक जंतु,कृंतक प्राणी,कृंतक जीव	एक प्रकार के छोटे जंतु जिनके मुँह में, विशेषकर कुतरने में सहायक, छोटे और पैने दाँत होते हैं	चुहिया,चूही,मुसटी,मूषिका	मादा चूहा
[being, beingness, existence]	the state or fact of existing	अस्तित्व,मौजूदगी,वज़ूद,वजूद,संभूति,विद्यमानता,सत्ता,हस्ती,भव,अस्ति,नमोनिशान	सत्ता का भाव	तत्त्व,तत्व,भूत,सत्त्व,सत्व,मूल द्रव्य	जगत का मूल कारण
[hand tool]	a tool used with workers' hands	हस्तोपकरण,हस्त उपकरण,हस्त-उपकरण	वे उपकरण जो हाथ द्वारा प्रयोग किए जाते हैं	हत्था,हाथा,हथेरा	खेत में पानी उलीचने का काठ का एक उपकरण

Figure 9.5: Excel file for hypernymy validation between UKC and Hindi



words_a_malayalam	gloss_a_malayalam	words_b_malayalam	gloss_b_malayalam
പ്രവർത്തനം	ഒരു കാര്യം നടക്കുന്ന അല്ലെങ്കില് ചെയ്യപ്പെടുന്നതായ ഭാവം	പതനം, നാശം, കീഴടങ്ങല്	യുദ്ധ സമയത്തു കൊട്ടാരം, പട്ടണം മുതലായവ തന്റെ കൈയിൽ നിന്നു മറ്റുള്ളവരുടെ കൈയിലേക്കു പോകുന്ന ക്രിയ.
വസ്തുവിന്റെ_ഭാഗം	ഒരു വസ്തുവിന്റെ ഏതെങ്കിലും ഭാഗം.	മുളങ്കണ്ണ്, പൊടിപ്പ്, നാളം	മുള പൊട്ടുന്ന സ്ഥാനം.
രണ്ടു_കാശ്	മണിയുടെ രൂപത്തിൽ ഉള്ള	കൈക്കുമ്പിൾ, കരസമ്പുടം.	രണ്ടു കൈകളും കൂട്ടിച്ചേർത്ത് ഉണ്ടാകുന്ന കുമ്പിളില് എന്തെങ്കിലും കൊടുക്കുകയോ വാങ്ങുകയോ ചെയ്യാം.
ചീത്തസമയം, അഴുക്കസമ	ചീത്തസമയം	ക്ഷാമ_കാലം	വളരെ കഷ്ടത്തോടുകൂടി അരി കിട്ടിയിരുന്ന കാലം.

Figure 9.6: Excel file for hypernymy validation within Malayalam

The validation steps are straightforward: first, the validator has to understand the UKC concept and use its gloss to understand its real meaning. Then she has to validate if the corresponding Hindi synset follows a synonym or hypernym relation. Figure 9.2 and Figure 9.3 show the flow diagrams for the validation tasks.

Then we will evaluate the validation results, which are the same excel files (refer to Figure 9.3, Figure 9.4, and Figure 9.5) with updated validator's responses. We use these files to estimate the accuracy of the mappings in the next step.

9.2.4 Accuracy Estimation

Accuracy refers to how close our observed sample is to the actual values in a larger population. Our sample of 1000 one-to-one mapped synsets was validated in the previous section, and in this step, we observed how many of the mappings were correct. Hence we estimate the accuracy of equivalence mappings between common concepts in Hindi.

We estimated the accuracy of 93 percent of equivalence mappings in Hindi. The 7 percent difference could be due to the expert' s knowledge of the sample' s concepts. We suggest validating with many validators and estimating the accuracy using a major voting approach.

9.2.5 Merging

This section describes integrating the input resource depending on the previous step' s validation. Estimating accuracy means that **x** percent of equivalent mappings are correct, and the mappings can be reused across all languages in the multilingual resource. We can extend the corresponding mappings for other languages once we have estimated the accuracy. For all other languages, we expand the mapping of synsets to concepts in the UKC.

Hindi is the reference resource to link other languages in IWN, and we used the same principle for merging the multilingual resource. We estimated that around 11283 synsets are mapped to the concepts in the UKC. Applying the same technique, we expand this mapping to other languages. For example, for Assamese, around 7314 concepts are mapped. Furthermore, all the languages are connected to language-independent concepts similar to the structure of the UKC.

9.3 Results

In this stage of methodology, we extract synsets from a synset-oriented resource and identify the corresponding concepts with the help of a language expert. Then we integrated the concepts with other languages to create a concept-oriented resource. IndoUKC is the name given to the resulting concept-oriented resource. After deleting duplicates and errored records, the IndoUKC resource has the first group of synsets in this phase. The IndoUKC concepts transfer to the UKC lexical concepts in the same way. The table below illustrates the number of concepts in phase II of the IndoUKC.

We observed from Table 9.1 that the concepts in each language vary, even the languages from the same language family. For example, Sanskrit and Nepali are from the same language family as Hindi. However, Sanskrit only has 19 percent coverage of common concepts from the total of 38070, while Nepali has 53 percent of concept coverage from 11713 synsets.

We prepared the concept coverage in the languages in the IndoUKC compared to Hindi (refer to Figure 9.7). The graph shows that Malayalam and Tamil have around 86 percent coverage compared to Hindi. However, Malayalam and Tamil belong to the Dravidian language family. We have one assumption that while preparing the wordnets from the Dravidian language family in the IWN project, the research team employed English wordnet to map the synsets.

Table 9.1: Status of IndoUKC concepts in phase II

No	Languages	#IWN synsets	#IndoUKC concepts
1	Assamese	14958	7314
2	Bengali	36346	9428
3	Bodo	15785	7346
4	Gujarati	35599	9342
5	Hindi	40371	11283
6	Kannada	22042	7880
7	Kashmiri	29469	8544
8	Konkani	32370	9218
9	Malayalam	30140	9719
10	Manipuri	16351	7328
11	Marathi	32721	8691
12	Nepali	11713	6218
13	Oriya	35284	9281
14	Punjabi	32364	9123
15	Sanskrit	38070	7213
16	Tamil	25419	9745
17	Telugu	21091	7633
18	Urdu	34280	9205

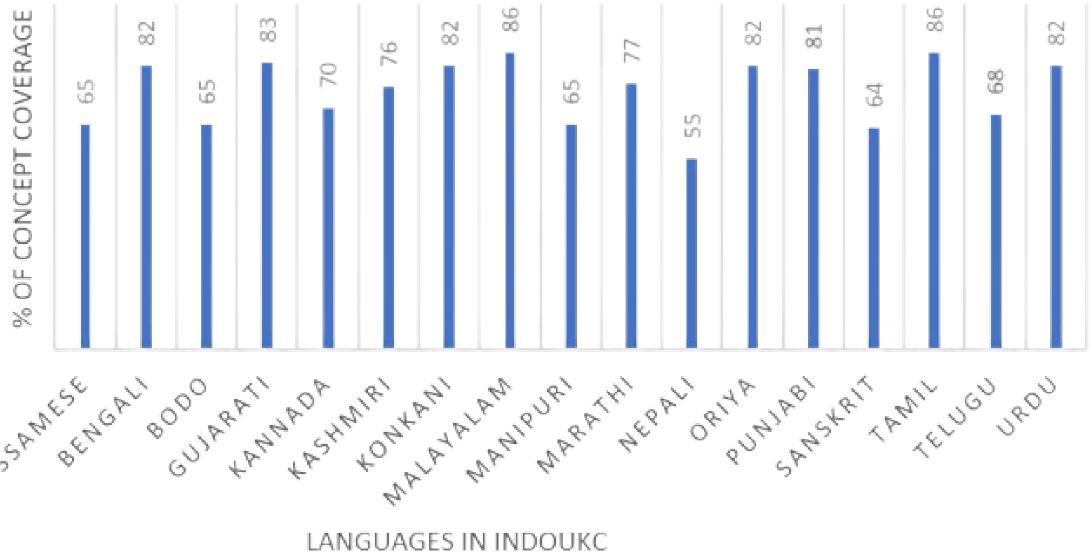


Figure 9.7: Concept coverage of languages VS Hindi

9.4 Summary

The chapter discussed how IWN concepts were merged. IndoUKC is the outcome resource of Indian languages with concepts. We described the five processes involved in the development of IndoUKC. The validation of the mappings is a crucial step in this chapter, and it was carried out utilizing excel sheets. The snapshots of the excel files were given so that the community could reuse them. We listed the concepts of the produced resource, which follows the UKC principle.

Chapter 10

Phase III: Merge One-to-Many Concepts

10.1 Introduction

The chapter aims to merge the one-to-many mapped synsets from the synset-oriented resource. For example, the concept "drum" in English is mapped to "mridhanga," "tabala," "nagaada," and "dhola." This example shows that the diversity of the Hindi language makes the resource create one-to-many mappings. Our proposed methodology aims to develop a resource that is not partial toward any language. Hence, we hope to represent the correct map of the "drum" and the four concepts in Hindi. If the language does not lexicalize the concept, IndoUKC will identify them as lexical gaps. The IndoUKC in this phase will contain more concepts than in the previous phase, and lexical gaps will be identified. The presence of lexical gaps helps us show how much a language is diverse.

The phase input is group B synsets from the synset-oriented resource. We give the synsets to the language experts to identify the correct match and move them to group C when the expert thinks it is an incorrect match or needs to learn the concept. Figure 10.1 shows the overview of phase III. This phase is about using an expert to identify the correct mappings.

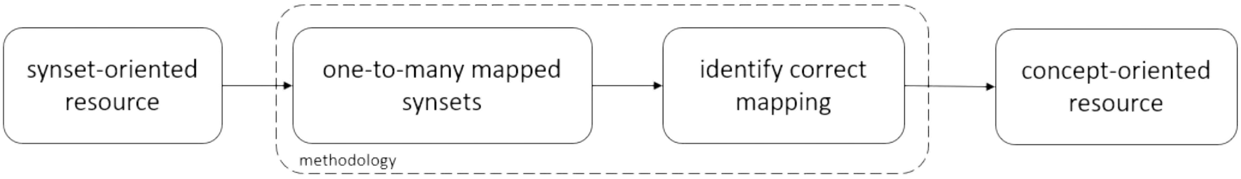


Figure 10.1: Overview of Phase III

10.2 The Process

The step will be merging the one-to-many mapped synsets from the synset-oriented resource. We grouped synsets into group B synsets with one-to-many mappings in chapter 8. Similar to the previous phase, we have multiple steps for merging the synsets and updating the concepts in the IndoUKC. Initially, we have to select one language to make the process faster. In our case, we have decided to select the language based on the availability of the language expert. In the second step, we ask the expert to identify the correct mappings out of many mappings. Our proposed methodology repeated the previously mentioned two steps to ensure our selected mappings so that we can extend to other languages. Figure 10.2 shows the process steps in phase III, and the detailed steps will be explained in the coming sections.

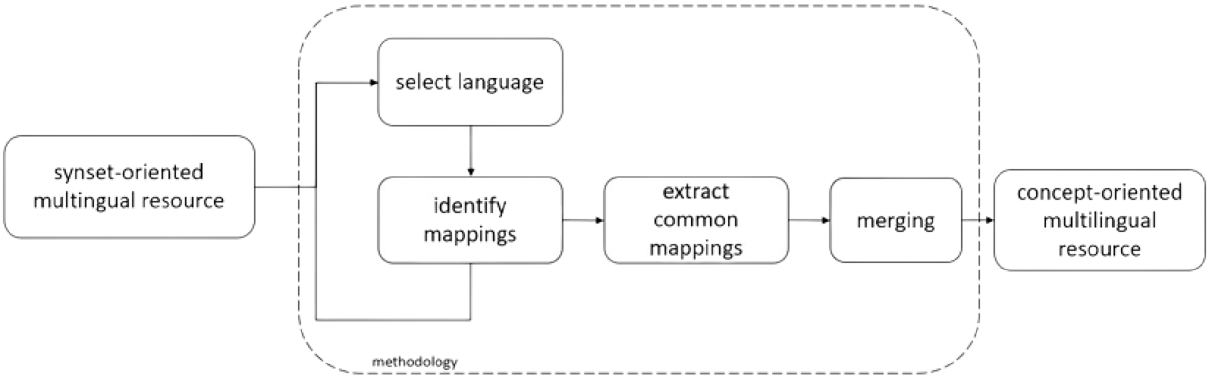

select language
synset-oriented multilingual resource
identify mappings
extract common mappings
merging
concept-oriented multilingual resource
methodology

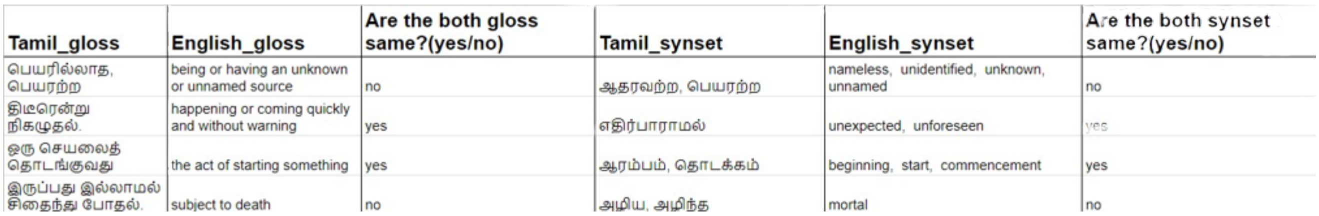
Figure 10.2: Phase III: Merge One-to-Many Concepts

10.2.1 Language Selection

In the language selection, we select one language with group B synsets. In the primary stage of the experiment, we decided to select languages based on the availability of language experts. The experts were native speakers of the languages between the ages of 28-56 and with minimum qualifications of a Master ’ s degree. Moreover, the languages we selected were from the same language family. Hence the synsets from the languages represented the same concepts.

10.2.2 Mappings Identification

This step aims to identify the correct mappings from the provided one-to-many mapped synsets. The input files for this step contain 6190 mappings in English and Malayalam and 5666 mappings in English and Tamil. The snapshot of the excel file we used for the Tamil experiment is given in Figure 10.3. The expert must respond true if the gloss and synset are the same; otherwise, false. The order we expected to do is first the gloss and then the synset. Because the expert first understands the concept from the gloss in both languages and then checks the words used in the languages.



Tamil_gloss	English_gloss	Are the both gloss same?(yes/no)	Tamil_synset	English_synset	Are the both synset same?(yes/no)
பெயரில்லாத, பெயரற்ற	being or having an unknown or unnamed source	no	ஆதரவற்ற, பெயரற்ற	nameless, unidentified, unknown, unnamed	no
திடீரென்று நிகழுதல்.	happening or coming quickly and without warning	yes	எதிர்பாராமல்	unexpected, unforeseen	yes
ஒரு செயலைத் தொடங்குவது	the act of starting something	yes	ஆரம்பம், தொடக்கம்	beginning, start, commencement	yes
இருப்பது இல்லாமல் சிதைந்து போதல்.	subject to death	no	அழிய, அழிந்த	mortal	no

Figure 10.3: Excel file for correct mappings identification

10.2.3 Extract Common Mappings

In this step, we need to finalize the mappings of the diverse concepts. Hence we run the experiment in more than one language. We collect the responses in excel files. Then we check the languages' responses to extract the common responses for the mappings. The extraction of mappings is effective if we identify the mappings for languages from the same language family since we hope the mapped synsets represent the same concepts.

We experimented with Malayalam and Tamil, both from the Dravidian language family. In this phase, we did not take a sample of the mapped synsets since the number of mappings was varied.

10.2.4 Merging

We used a sample set of 1000 selected mappings from the Malayalam and Tamil languages. We computed the interrater agreement using the percent agreement. The raters from both languages agree, so there is a 100 percent agreement. We calculated another inter-rater agreement, kappa, where we subtract the estimated level of chance agreement from the observed level of agreement, dividing by the maximum possible non-chance agreement. The observed agreement is one, and the chance agreement is 0.637.

Furthermore, which eventually indicates that raters agree 100 percent. Based on the agreement score, we now have synsets from Indian languages mapped to the UKC' s universal lexical concepts. Hence we extend the mappings of synsets for other languages. So the current phase, the concept-oriented resource will have more concepts and can be interconnected to any language part of the UKC.

10.3 Results

Table 10.1 shows the status of the IndoUKC in phase III. An average of 9K concepts are available in each language in IndoUKC. Ten languages in the IndoUKC is more than 10K, which will significantly support improving the quality of NLP applications in Indian languages. The total number of

Table 10.1: Status of IndoUKC concepts in phase III

No	Languages	#IWN synsets	#IndoUKC concepts
1	Assamese	14958	9093
2	Bengali	36346	11119
3	Bodo	15785	9250
4	Gujarati	35599	11104
5	Hindi	40371	11737
6	Kannada	22042	9956
7	Kashmiri	29469	10423
8	Konkani	32370	10938
9	Malayalam	30140	11184
10	Manipuri	16351	9137
11	Marathi	32721	10379
12	Nepali	11713	7093
13	Oriya	35284	10927
14	Punjabi	32364	10764
15	Sanskrit	38070	8710
16	Tamil	25419	9884
17	Telugu	21091	9737
18	Urdu	34280	10864

synsets in IWN includes all errored records, synsets as named entities, and repeated entries of synsets, so the comparison of the number of synsets and the number of concepts is inappropriate.

Figure 10.4 shows the number of group A synsets before and after phase III. After phase III, there is a noticeable difference in the increase of common concepts for all languages. It is to be noted that even if the language we selected for finding the equivalent mappings were from the Dravidian language family, languages from other language families also increased. We plan to experiment with more languages to validate the similarity of concepts across the languages, even from different linguistic families.

Moreover, only some group B synsets or one-to-many mapped synsets exist. After phase III, the group B synsets are split into groups A and C. In the next chapter, we will explain how we merge the group C synsets.

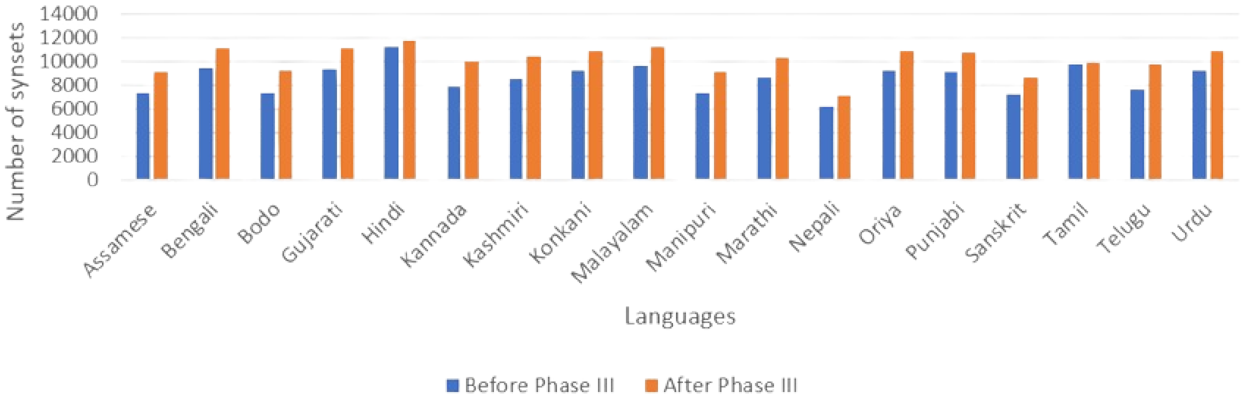


Figure 10.4: Number of group A synsets before and after phase III

10.4 Summary

The chapter explained the steps in phase III of the proposed methodology for merging the one-to-many concepts from languages. This chapter is an important phase of the resultant concept-oriented resource generation since we hope to develop our resources as language-independent or diversity-aware. When we can not find the correct concept corresponding to the synset, we do not link it to a general or specific concept; instead, we indicate it as a lexical gap for the language. Hence we can ensure the resultant resource is diversity-aware.

Chapter 11

Phase IV : Merge Language-specific Concepts

11.1 Introduction

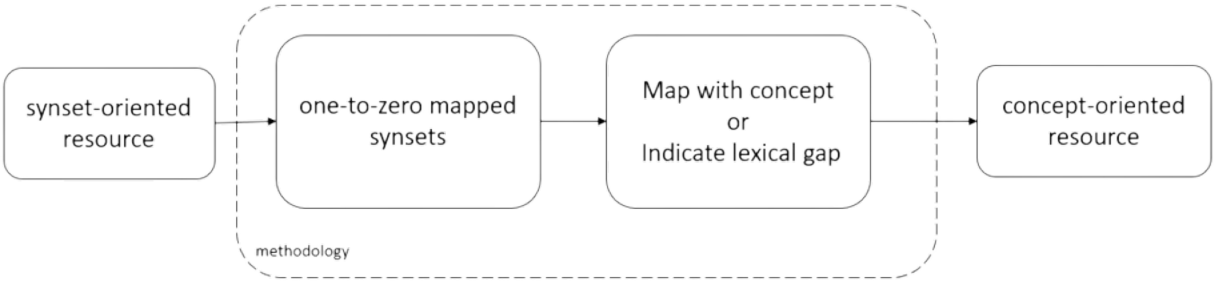

synset-oriented resource
one-to-zero mapped synsets
Map with concept or Indicate lexical gap
concept-oriented resource
methodology

Figure 11.1: Overview of the Phase IV

This chapter brings together the language-specific concepts found in the synset-oriented resource. We examine the one-to-zero mappings from the group C synsets and show how to determine the concept that corresponds to them. There are no English synsets for the synsets in group C in IWN; this part of the process ensures that equivalence mappings or lexical gaps are identified. According to the thesis, the group C synset is a language-specific concept that cannot be expressed lexically. Nonetheless, we assume the lack of a corresponding synset is because IWN synsets need to be fully mapped to all PWN. We can consider a standardized approach for adding a concept to a concept-oriented resource and representing concepts from any language in the final stage. Figure 11.1 shows the overview of phase IV.

The proposed methodology utilizes the synset-oriented resource as an input and incorporate the concepts into the IndoUKC, a concept-oriented resource that was built in the previous step. This stage differs from the previous one in that we must locate the concepts. If a concept is not found in English, it is added to the list of universal lexical concepts. Then mention that the concept has a lexical gap in English. Figure 11.3 shows the flow diagram of phase IV.

11.2 The Process

The steps for integrating the language-specific concepts are explained in this section (refer to Figure 11.2). To demonstrate how to implement the merging of the group C synsets, we start by choosing a language. We decided to sample the synsets further because the number of group C synsets is more extensive in all languages of the IWN. The synsets are sampled based on the domain to which they belong. Food, plants, and kinships are examples of domains. As a result, we will demonstrate how to include a domain of concepts in the IndoUKC.

11.2.1 Language Selection

The language we select depends on the expert. Expert has to be good knowledge of English since she needs to translate the synsets into English and correctly map the concept according to the meaning of the synset. Hence, we experimented with the process with a graduate from the age

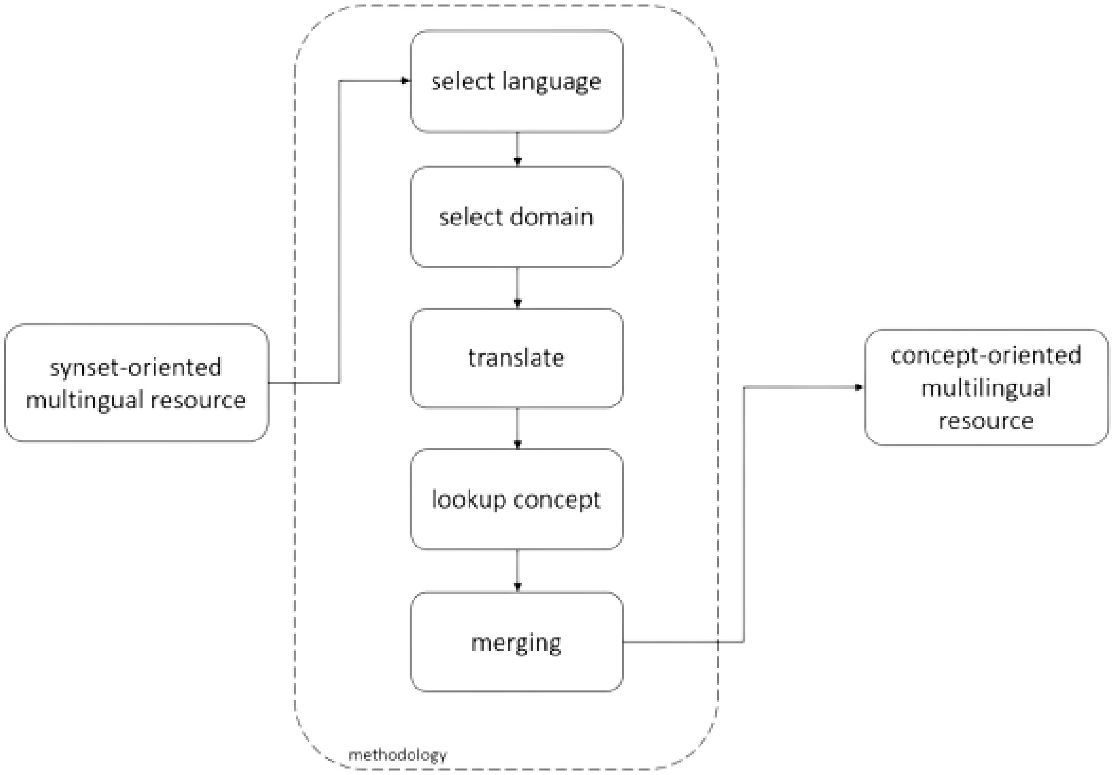

select language
select domain
translate
lookup concept
merging
synset-oriented multingual resource
concept-oriented multilingual resource
methodology

Figure 11.2: Phase IV : Merge Language-specific Concepts

group of 28-56.

We selected Malayalam and have 18956 group C synsets after phase III. When we fixed the group B synsets, we had the synsets that required more attention from the expert. Hence, we shifted the synsets into group C.

11.2.2 Domain Selection

Going through each synset for a single language will take the expert a lot of time and effort. As a result, we divide the resource into smaller groups, as shown in Table 11.1. We asked the expert to go through a random sample of 1000 group C synsets and categorize the synsets based on the domain to create a small group.

Experts in the Malayalam language examined over 1000 synsets and classified them into 20 groups, finding around 300 synsets that were re-

Table 11.1: Domains in sample set of 1000 group C synsets

Domain	#synsets	Domain	#synsets
animals	18	astrology	20
birds	58	buildings	5
clothing	14	cultural	92
diseases	26	family	17
farming	6	food	29
human body	19	linguistics	27
literature	40	measurements	28
music	29	ornaments	16
plants	104	festivals	30
sports	15	tools	55

peated entries. Table 11.1 lists the domain names and the number of synsets in each group. We discovered that the group C synsets are distributed across many language domains rather than restricted to one or two. As a result, utilizing the phase II method, we can link more languages of the IWN by adding concepts corresponding to the group C synsets of one language.

11.2.3 Translation

We will proceed to the next phase by choosing a domain in which the expert is confident because she must comprehend the concept she is translating. Based on the source synset lemmas, gloss, and examples, we first requested the expert confirm whether the concept represents a lexical gap in the target language.

When a lexical gap is identified, it signifies that the concept is either unknown in the target language's culture or can only be lexicalized using word combinations. The expert must provide a rough gloss of the target language's concept. Approximating the meaning and identifying a more generic meaning idea in the source language instead of lexical gap identification is not a correct solution because it violates the IndoUKC resource's diversity feature.

The translation process is organized as follows if the expert identified the concept as not a lexical gap: after understanding the synset, begin with the gloss translation, then translate the synsets. The gloss translation should provide a clear description of the concept. Meanwhile, the writing style and manner should be at least as good as the gloss in the source language. The synset lemmas must then be translated, which is unlikely to be a one-to-one mapping between the source and target languages. We instructed the expert to use authorized dictionaries to collect synsets in the target language.

Figure 11.4 shows the excel file snippet we used to align the concept from the domain "farming." Out of the sample of 1000 synsets, we found six synsets related to the "farming" domain—two of the synsets aligned with the concepts from the universal lexical concepts.

Figure 11.5 shows the excel file snippet we used for concepts that are lexical gaps from the domain "farming." We identified four "farming" domain concepts as lexical gaps in English.

11.2.4 Concept Lookup

In this step, we check if the concept we want to add already exists in the list of universal lexical concepts. Otherwise, we duplicate the concepts. The step of concept lookup was performed using the UKC' s web application for lexicon search[3], which will be explained in chapter 13. We use the translated synsets in the lexicon search service to find the suitable concept.

11.2.5 Merging

There are two possibilities to consider: when the concept can be translated and when it cannot. With the help of our experts, we found the relevant concept of the group C synset in the previous concept lookup phase. The group C synset is then aligned. We detect the concept as a lexical gap and apply the gloss defined in the Translation step when the group C synset is not translatable.

11.3 Summary

The final phase of the proposed methodology was to merge a synset-oriented resource and help to add new concepts to the IndoUKC. The group C synsets from the IWN were handled in this phase. The experts helped us understand the complexity of merging group C. Group C synsets have concepts from diverse domains that still need to be uncovered in other languages. This chapter provided much scope for future research and evolved the Indian languages in the culture-specific domains.

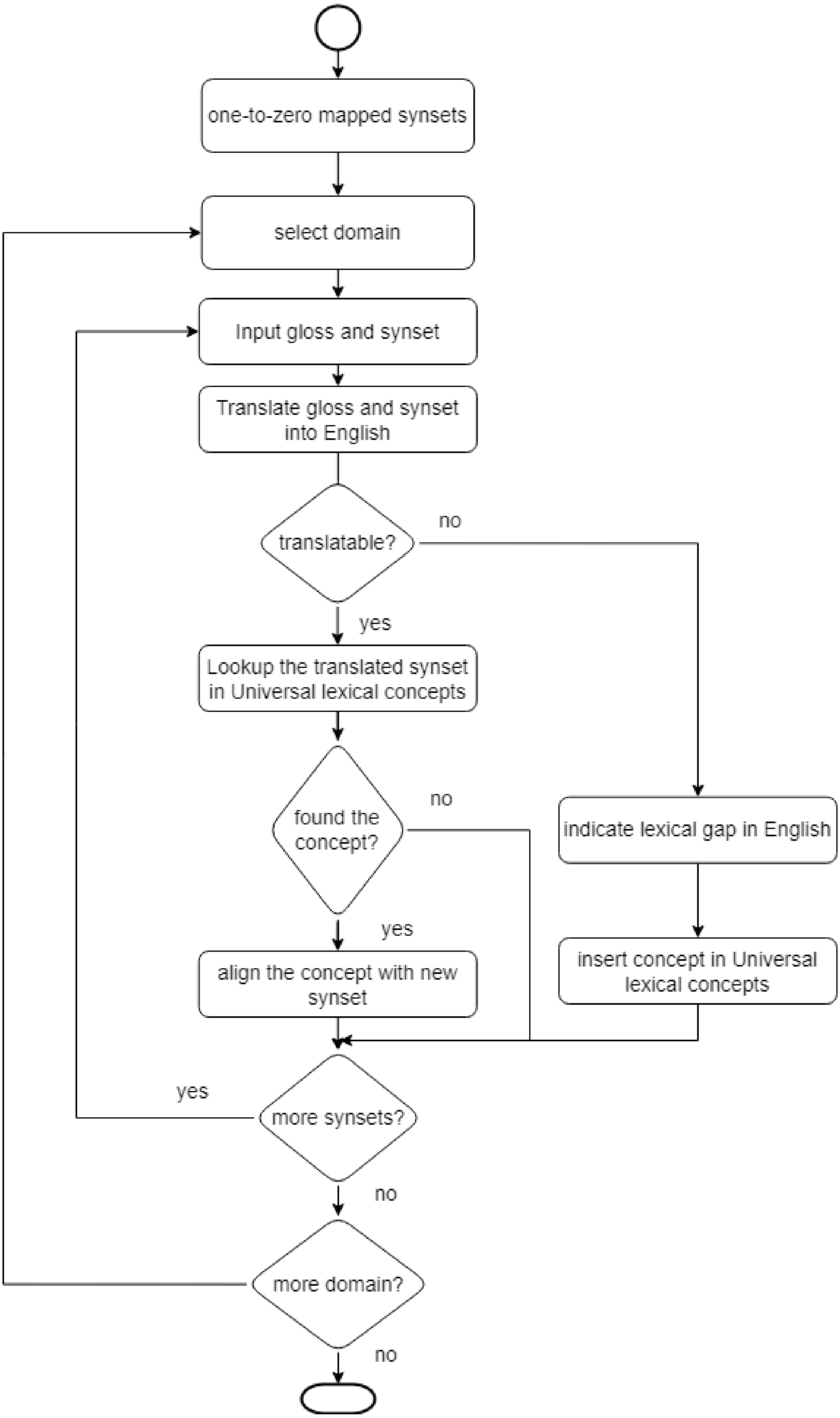
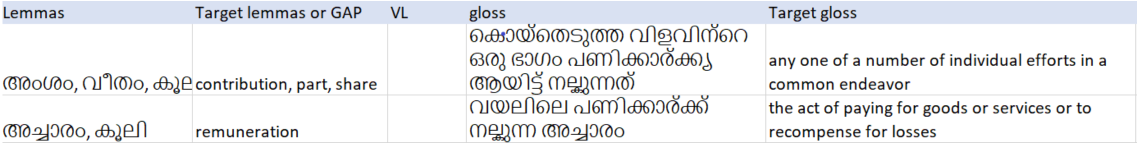
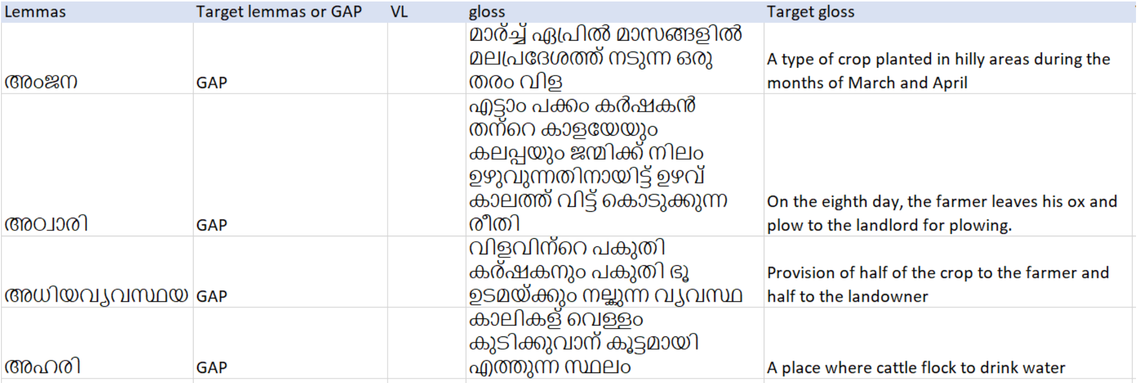

one-to-zero mapped synsets
select domain
Input gloss and synset
Translate gloss and synset into English
translatable?
no
yes
Lookup the translated synset in Universal lexical concepts
found the concept?
no
yes
indicate lexical gap in English
insert concept in Universal lexical concepts
align the concept with new synset
yes
more synsets?
no
more domain?
no

Figure 11.3: Flow diagram of Phase IV



Lemmas	Target lemmas or GAP	VL	gloss	Target gloss
അംശം, വീതം, കൂല	contribution, part, share		കൊയ്തെടുത്ത വിളവിന്റെ ഒരു ഭാഗം പണിക്കാര്ക്ക്യ ആയിട്ട് നല്കുന്നത്	any one of a number of individual efforts in a common endeavor
അച്ചാരം, കൂലി	remuneration		വയലിലെ പണിക്കാര്ക്ക് നല്കുന്ന അച്ചാരം	the act of paying for goods or services or to recompense for losses

Figure 11.4: Example of mapping concept from Farming domain



Lemmas	Target lemmas or GAP	VL	gloss	Target gloss
അംജന	GAP		മാര്ച്ച് ഏപ്രിൽ മാസങ്ങളിൽ മലപ്രദേശത്ത് നടുന്ന ഒരു തരം വിള	A type of crop planted in hilly areas during the months of March and April
അഠാരി	GAP		എട്ടാം പക്കം കർഷകൻ തന്റെ കാളയേയും കലപ്പയും ജന്മിക്ക് നിലം ഉഴുവുന്നതിനായിട്ട് ഉഴവ് കാലത്ത് വിട്ട് കൊടുക്കുന്ന രീതി	On the eighth day, the farmer leaves his ox and plow to the landlord for plowing.
അധിയവ്യവസ്ഥയ	GAP		വിളവിന്റെ പകുതി കര്ഷകനും പകുതി ഭൂ ഉടമയ്ക്കും നല്കുന്ന വ്യവസ്ഥ	Provision of half of the crop to the farmer and half to the landowner
അഹരി	GAP		കാലികള് വെള്ളം കുടിക്കുവാന് കൂട്ടമായി എത്തുന്ന സ്ഥലം	A place where cattle flock to drink water

Figure 11.5: Example of Lexical Gap identification from Farming domain

Part IV

Evaluation

Chapter 12

IndoUKC

12.1 Introduction

While the possibility of a supra-lingual representation of lexical meaning underlies all multilingual lexical resources presented in this thesis, only the IndoUKC makes an explicit commitment in this direction, representing the concept hierarchy separately from the individual languages [3]. The IndoUKC does not believe a unified concept graph can adequately capture all languages' lexical meanings. The capacity to express word meanings and their hierarchy on both the supra-lingual and language-specific levels, the former using concepts and the latter using synsets, is a major distinguishing characteristic compared to all previously described multilingual resources.

Figure 12.1 shows the IndoUKC representing the concept and lexical gaps in English, Malayalam, and Hindi. From Figure 12.1, we can see that the concepts in languages have one-to-one mappings. For example, "relative" in English is lexicalized as "ബന്ധു " (bandhu) in Malayalam and "रिश्तेदार "(rishthedhar) in Hindi. IndoUKC is capable of representing lexical gaps in any language. As shown in Figure 12.1, the concept of "cousin" in English is a lexical gap between Malayalam and Hindi. Also, "मौसेरा बाई" (mousera bhai) in Hindi is a lexical gap in English.

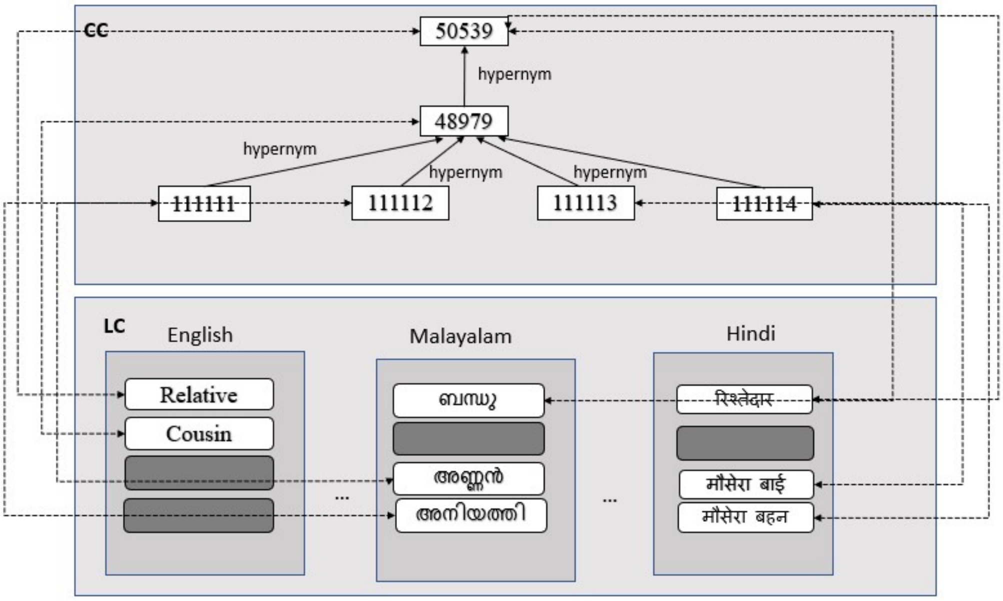


Figure 12.1: IndoUKC

12.2 Incompleteness

We consider lexical elements of the language as one factor that decides the quality of the resource. The languages we used in our study still need to be completed. That means the concepts need to be added due to the lack of progressive development. In addition, cross-lingual lexical alignments are another factor that affects the quality of the multilingual resource, which means that only correct associations between concepts and synsets are. No matter how advanced a language becomes, it will always be devoid of many words and reflect the misunderstandings and errors of those who created it. We employ a quantitative metric to assess a language's quality.

Language Incompleteness is a concept that has been offered. **LanInc(l)** is a straightforward extension of the idea of the incompleteness of logical languages and theories, with its counterpart notion of Language Coverage **LanCov(l)**. The goal is to take advantage of the fact that the CC can be used to tally how much of a language's domain of interpretation, defined as a set of synsets, is not lexicalized by that language (a computational representation).

$$AbsLanCov(l) = |concepts|$$
$$LanCov(l) = \frac{|AbsLanCov(l)|}{|concepts(UKC)| - |Gaps(l)|}$$

$$LanInc(l) = 1 - LanCov(l)$$

where concepts(l) refers to the set of concepts lexicalized by a language l, concepts(UKC) refers to the concepts in the UKC, and Gaps(l) refers to the lexical gaps in l. The Absolute Language Coverage is abbreviated as AbsLanCov(l). In Table 12.1, we compute the LanInc(l) for IndoUKC. We modified the formula according to IndoUKC. We are considering Gaps(l) as empty in the initial stage of the study.

$$AbsLanCov(l) = |synsets|$$
$$LanCov(l) = \frac{|AbsLanCov(l)|}{|concepts(IndoUKC)|}$$

The measurement emphasizes the need to continue to evolve languages and identifies which languages require additional attention in future efforts, like finding the lexical gaps and increasing the coverage.

12.3 Results

We evaluated the improvement of mappings of meanings across 14 Indian languages. We used English as a reference language for creating the mappings. We used 77 noun concepts to sample the number of mappings between English and Indian languages. We selected concepts tied to common mappings across all languages that are not unique to Indian languages to

Table 12.1: Language Coverage and Language Incompleteness of IndoUKC

Languages	AbsLanCov(l)	LanCov(l)	LanInc(l)
Assamese	14954	0.37	0.67
Bengali	36333	0.89	0.11
Bodo	15781	0.39	0.61
Gujarati	35570	0.87	0.13
Hindi	39239	0.96	0.04
Kannada	22027	0.54	0.46
Kashmiri	29441	0.72	0.28
Konkani	32356	0.79	0.21
Malayalam	29047	0.71	0.29
Manipuri	16313	0.40	0.60
Marathi	29700	0.73	0.27
Nepali	11659	0.29	0.71
Oriya	35275	0.86	0.14
Punjabi	32336	0.79	0.21
Sanskrit	23117	0.57	0.43
Tamil	25417	0.62	0.38
Telugu	21087	0.52	0.48
Urdu	34105	0.83	0.17

avoid the partiality of one language within the evaluation. We observed a pattern while sampling mappings between English and Indian languages: the number of mappings varies for languages. Figure 12.2 shows the number of mappings in each language.

We used semantic similarity as a feature to evaluate the proposed resource. Semantic similarity estimates the semantic closeness of senses from two different languages. We calculated semantic similarity for English and Indian language glosses converted into vectors using BERT sentence transformers. We clustered semantic scores depending on each language' s overall average of 0.4. Hence we considered mapping greater than 0.4 as the correct association. Figure 12.3 shows a semantic similarity comparison be-

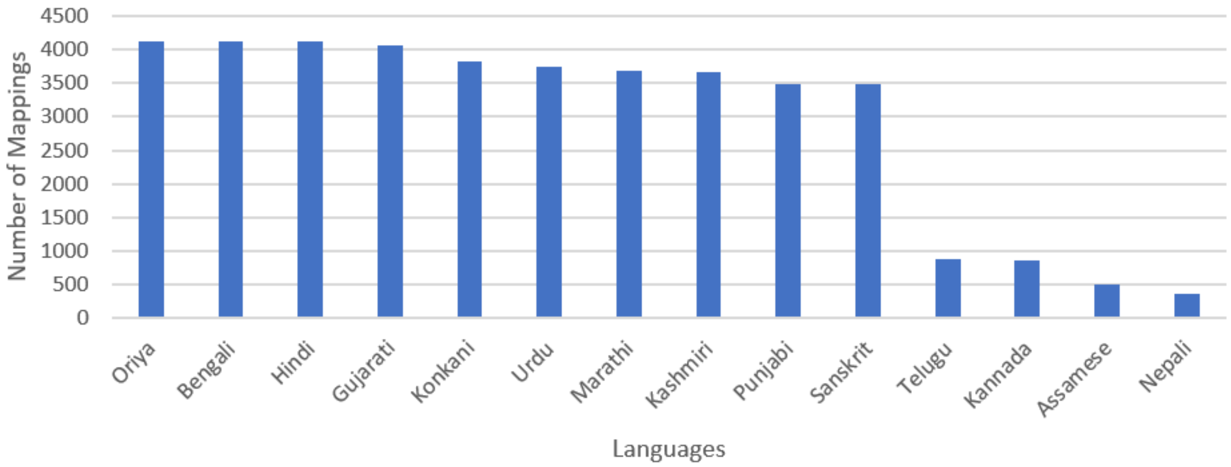


Figure 12.2: Number of mappings for each language

tween IWN and IndoUKC. The evaluation showed that the final outcome, IndoUKC is different from IWN and our methodology helped to improve the cross-lingual associations.

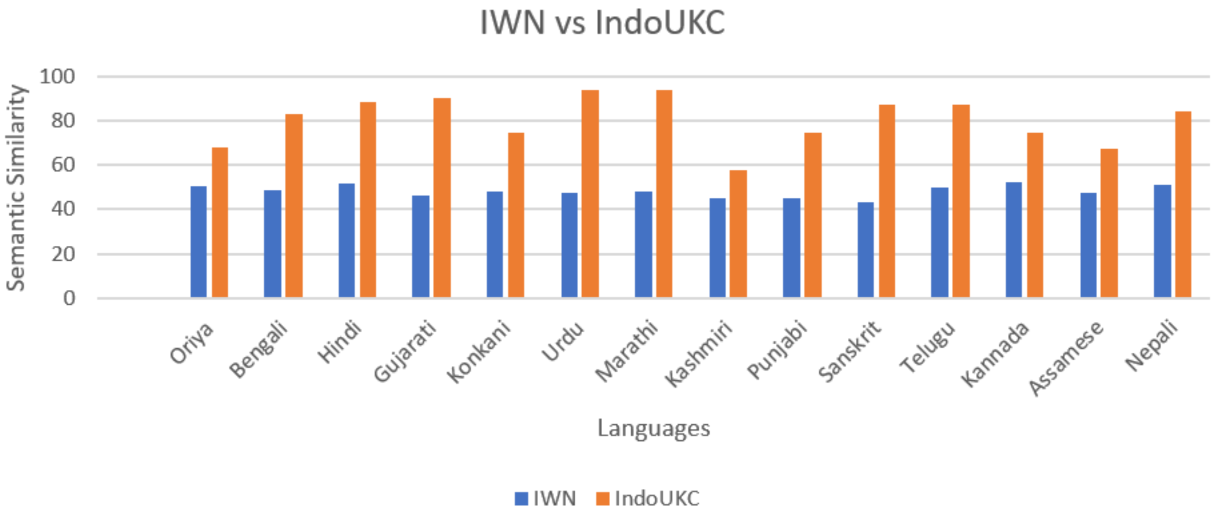


Figure 12.3: Mappings comparison: IndoWordNet vs. IndoUKC

12.4 Summary

The chapter evaluates the resource, IndoUKC, for the mapping ability compared to other multilingual resources. The chapter showed that the similarity of meanings across languages is an important parameter that still needs improvement for IndoUKC as it progresses.

Part V

Implementation

Chapter 13

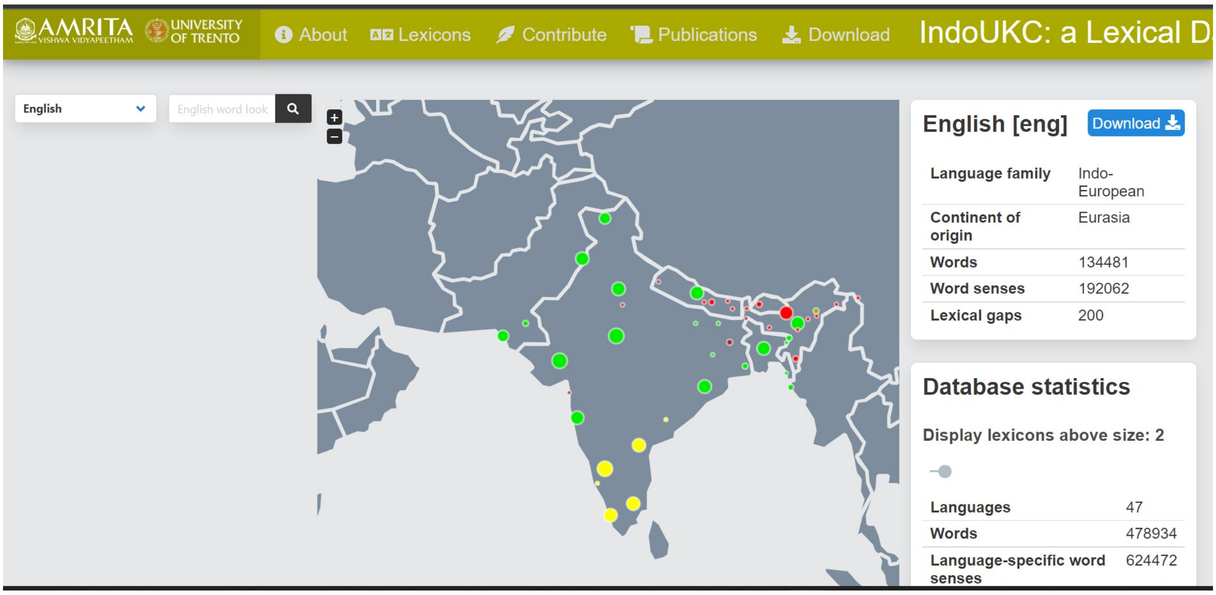
IndoUKC Website

13.1 Introduction

The chapter provides the details of the IndoUKC resource published online [3]. This resource aims to improve the coverage of the language and use it as a knowledge base for finding language-specific concepts. We show here the snapshots of the website and available services the website offers currently. We hope to update the website versions with more coverage of lexical items and languages in the future. Moreover, the website will also act as a platform for future collaboration. We hope to support linguists and researchers in improving NLP research using our resource IndoUKC.

13.2 IndoUKC website

The website is published in the link, http://indo.ukc.datascientia.eu/concept. The website has an about page that describes the project and related works (refer to Figure 13.2. The lexicons tab of the website shows the services provided with the IndoUKC resource. We provided a contribute tab to collect the collaboration input from others. We list our publications in the publications tab. We also have a download tab redirecting to our group' s page for datasets where users can download a resource. The snapshot



English [eng]

Language family	Indo-European
Continent of origin	Eurasia
Words	134481
Word senses	192062
Lexical gaps	200

Database statistics

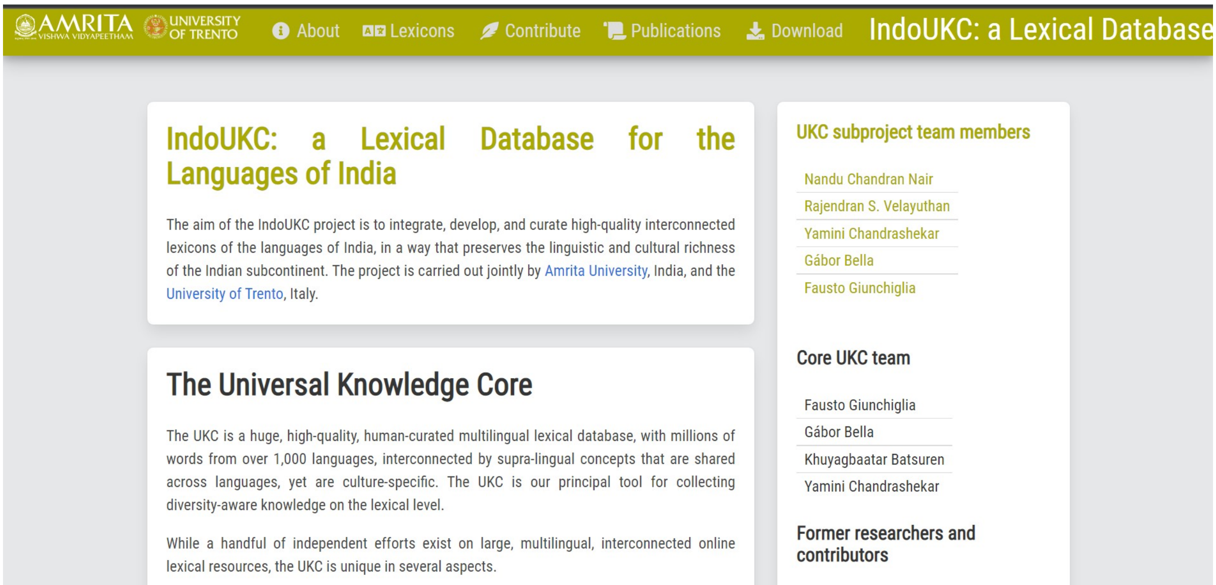
Display lexicons above size: 2

Languages	47
Words	478934
Language-specific word senses	624472

Figure 13.1: IndoUKC Website

of the website is shown in Figure 13.1. We used the opportunity at the conference LREC 2022 to present our work on IndoUKC and conducted a website demo in the conference's workshop section. We collected the feedback for our website to apply in our future version. One of the suggestions is to have images as part of the concept description to help clarify the concept.

The current version of the website has 47 languages from different parts of India (refer to the map shown in Figure 13.1). With our group' s team support, we managed to include the under-resourced languages and hope to include more in the future [3]. Languages we imported from other sources will also be imported here to enrich the culture-specific vocabularies from India. We hope to become one resource that can be used to study the language diversity and similarity across Indian languages.



IndoUKC: a Lexical Database for the Languages of India

The aim of the IndoUKC project is to integrate, develop, and curate high-quality interconnected lexicons of the languages of India, in a way that preserves the linguistic and cultural richness of the Indian subcontinent. The project is carried out jointly by Amrita University, India, and the University of Trento, Italy.

The Universal Knowledge Core

The UKC is a huge, high-quality, human-curated multilingual lexical database, with millions of words from over 1,000 languages, interconnected by supra-lingual concepts that are shared across languages, yet are culture-specific. The UKC is our principal tool for collecting diversity-aware knowledge on the lexical level.

While a handful of independent efforts exist on large, multilingual, interconnected online lexical resources, the UKC is unique in several aspects.

UKC subproject team members

Nandu Chandran Nair
Rajendran S. Velayuthan
Yamini Chandrashekar
Gábor Bella
Fausto Giunchiglia

Core UKC team

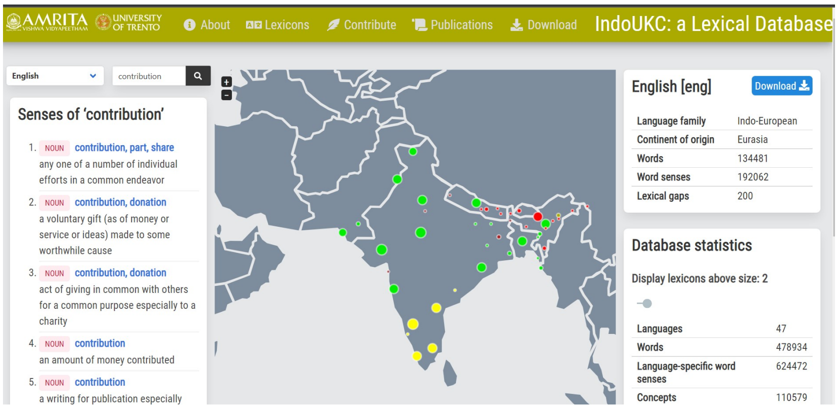
Fausto Giunchiglia
Gábor Bella
Khuyagbaatar Batsuren
Yamini Chandrashekar

Former researchers and contributors

Figure 13.2: IndoUKC Info Page

13.3 Lexicon Search

In the current version of the IndoUKC website, the service available is lexicon search. We can search a lexicon in the available 47 languages on the website. Figure 13.3 describes how we can look up a concept on the IndoUKC website. The search results will provide the list with different senses for the searched word. The user can choose the intended sense, like the one shown in Figure 13.4.



Senses of 'contribution'

1. NOUN contribution, part, share
   any one of a number of individual efforts in a common endeavor
2. NOUN contribution, donation
   a voluntary gift (as of money or service or ideas) made to some worthwhile cause
3. NOUN contribution, donation
   act of giving in common with others for a common purpose especially to a charity
4. NOUN contribution
   an amount of money contributed
5. NOUN contribution
   a writing for publication especially

English [eng]

Language family	Indo-European
Continent of origin	Eurasia
Words	134481
Word senses	192062
Lexical gaps	200

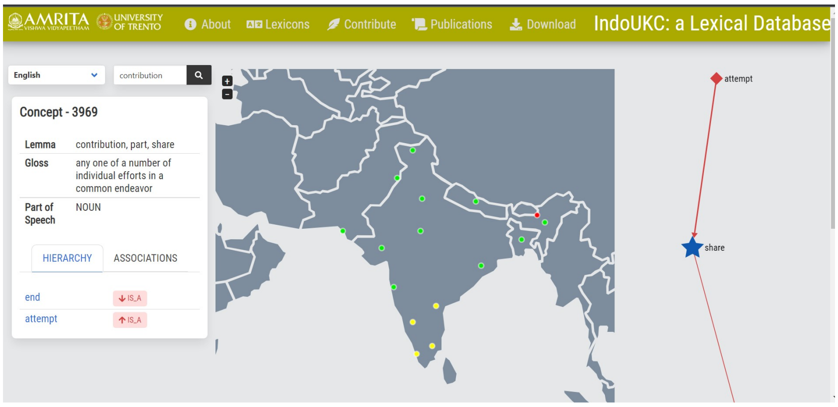
Database statistics

Display lexicons above size: 2

Languages	47
Words	478934
Language-specific word senses	624472
Concepts	110579

Figure 13.3: Lexicon search in IndoUKC Website

Figure 13.4 shows the search result for the concept "contribution." Figure 13.4 details the concept with the id of the concept, lemma, gloss of the concept, and part of speech. Moreover, the hierarchy and association details are also available. The page also provides a visual representation of the hierarchy. Once we select the concept from the list, we get the details of the concept, as shown in Figure 13.4.



English
contribution

Concept - 3969

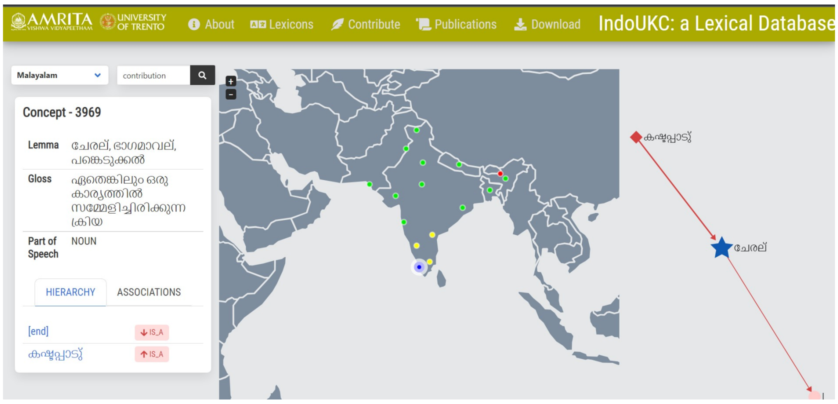
Lemma	contribution, part, share
Gloss	any one of a number of individual efforts in a common endeavor
Part of Speech	NOUN

HIERARCHY ASSOCIATIONS

end	↓ IS_A
attempt	↑ IS_A

Figure 13.4: Details of the concept

The IndoUKC lexicon search in Figure 13.3 shows is in English. The website allows changing the exact search in the available 47 languages just by changing the language in the drop-down list on the website. As shown in Figure 13.5, we show the search for the concept "contribution" in Malayalam.



Malayalam
contribution

Concept - 3969

Lemma	ചേരല്, ഭാഗമാവല്, പങ്കെടുക്കൽ
Gloss	ഏതെങ്കിലും ഒരു കാര്യത്തിൽ സമ്മേളിച്ചിരിക്കുന്ന ക്രിയ
Part of Speech	NOUN

HIERARCHY ASSOCIATIONS

[end]	↓ IS_A
കഷ്ടപ്പാടു്	↑ IS_A

Figure 13.5: Lexicon search in Malayalam

13.4 Summary

This chapter is dedicated to explaining the current version of IndoUKC website. We hope to have a progressive update of the different website versions with more coverage in languages and lexical elements. IndoUKC, like PWN, can be used for word-sense disambiguation, information retrieval, automatic text classification, automatic text summarization, machine translation, and even crossword problem development. Furthermore, we hope to make the IndoUKC resource a strong foundation for the NLP research of Indian languages.

Part VI

Conclusion and Future Work

Chapter 14

Conclusion

The thesis proposed a methodology for the lexical resource development community. Our methodology supports the development of under-resourced languages and enriches the resource with our proposed collaborative environment. Our proposed methodology integrated multiple methods from the state-of-the-art to generate a concept-oriented, language-independent, multilingual lexical resource. As a test case, we have implemented our approach for Indian languages to exhibit the importance of capturing the diversity and uniqueness of languages.

The thesis is structured to make the reader understand the problem and the difficulty of achieving the solution. The work presented here is not a single person’ s work but a team’ s work to preserve languages with diversity awareness. We provided the descriptions and structural details of existing lexical resources to understand the context of the study. Moreover, an explanation of existing resources also helps to understand the challenges in multilingual lexical resources. Furthermore, we showed why UKC is an excellent tool to use as part of our methodology.

The UKC is a multilingual lexical resource that captures lexical diversity, and we used its principle to achieve a diversity-aware feature in the final resource. Supporting lexical gaps and merging language-specific meanings into a central concept graph is crucial to reaching these long-term goals. Unlike IWN, the final resource, IndoUKC, is connected to other world languages, and a setup for collaboration is implemented for progressive development.

The proposed resource of Indian languages is available as planned in the initial research stage. IndoUKC is easily accessible online for searching lexicons and concepts. More importantly, our resource IndoUKC is different from IWN since IndoUKC is connected with around 1000 other world languages and gives importance to each language' s unique concepts. This study aims to motivate more languages to be part of such a community of language development and preserve languages as it is. With our concept-oriented resource, we hope to improve the performance of NLP applications in Indian languages. Moreover, reduce the digital language divide for Indian languages in coming future.

Chapter 15

Future Work

The proposed methodology showed the approach for adding new concepts that are common and language-specific. So the major challenge for the future is to complete the resource with more lexical items. Moreover, we hope to get feedback on our proposed approach in various use cases.

Generally, we would like to increase coverage in the languages starting with kinship terms. The domain of kinship is a hot topic, and we are interested in studying the similarity and diversity among 1000 languages. Unlike IWN, IndoUKC contains kinship domain concepts. In order to help the long-term preservation of the Sanskrit language, we also desire to broaden the list of concepts, such as Sanskrit lexicons from the Bhagavad Gita.

We developed an environment for collaboration with a service of lexicon search. However, we are hoping to show diversity-aware in languages with the help of multimedia to make the understanding of sense clear. Instead of having text form definitions for lexical gaps, we can use audio, image, or video to represent the sense. In this way, we could increase the coverage using crowdsourcing.

In order to add more concepts to Malayalam and other languages, we are hoping to collaborate with the research teams in India. We have started

Tamil as our subsequent use case because it belongs to the same language family as Malayalam, and language experts are readily available. Therefore, we use the conferences to network with people and ask for their help in contributing to new languages.

Bibliography

[1] Noryusliza Abdullah and Rosziati Ibrahim. Managing information by utilizing wordnet as the database for semantic search engine. *International Journal of Software Engineering and Its Applications*, 9(5):193–204, 2015.

[2] Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. Cognet: a large-scale cognate database. In *Proceedings of ACL 2019, Florence, Italy*, 2019.

[3] Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Ashgar Cheema, and Fausto Giunchiglia. Language diversity: Visible to humans, exploitable by machines. *arXiv preprint arXiv:2203.04723*, 2022.

[4] Francis Bond, Luis Morgado Da Costa, Michael Wayne Goodman, John Philip McCrae, and Ahti Lohk. Some issues with building a multilingual wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3189–3197, 2020.

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

[6] Sonja E Bosch and Marissa Griesel. Strategies for building wordnets for under-resourced languages: The case of african languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 38(1):1–12, 2017.

[7] Lyle Campbell. Ethnologue: Languages of the world, 2008.

[8] Steven CF Kui, Pui Keong Chow, Glenna So Ming Tong, Shiu-Lun Lai, Gang Cheng, Chi-Chung Kwok, Kam-Hung Low, Man Ying Ko, and Chi-Ming Che. Robust phosphorescent platinum (ii) complexes containing tetradentate o^ n^ c^ n ligands: Excimeric excited state and application in organic white-light-emitting diodes. *Chemistry–A European Journal*, 19(1):69–73, 2013.

[9] Debasri Chakrabarti and Pushpak Bhattacharyya. Creation of english and hindi verb hierarchies and their application to hindi wordnet building and english-hindi mt. In *Proceedings of the Second Global Wordnet Conference, Brno, Czech Republic*. Citeseer, 2004.

[10] Daniel Cunliffe, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope. Natural language processing for under-resourced languages: Developing a welsh natural language toolkit. *Computer Speech & Language*, 72:101311, 2022.

[11] Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D Pawar. *The WordNet in Indian Languages*. Springer, 2017.

[12] Nancy Dorian. Small-language fates and prospects: Lessons of persistence and change from endangered languages: Collected essays. In *Small-Language Fates and Prospects*. Brill, 2014.

[13] Christiane Fellbaum. Wordnet. *The Encyclopedia of Applied Linguistics*, 2012.

[14] Christiane Fellbaum and Piek Vossen. Connecting the universal to the specific: Towards the global grid. In *International Workshop on Intercultural Collaboration*, pages 1–16. Springer, 2007.

[15] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017, 2017.

[16] 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, 18-24 March 2018*, 2018.

[17] Fausto Giunchiglia and Mattia Fumagalli. Concepts as (recognition) abilities. In *FOIS*, pages 153–166, 2016.

[18] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: Hlt*, pages 771–779, 2008.

[19] Juliane House. Translation quality assessment: Past and present. In *Translation: A multidisciplinary approach*, pages 241–264. Springer, 2014.

[20] Michael Ibba and Dieter Söll. Quality control mechanisms during translation. *Science*, 286(5446):1893–1897, 1999.

[21] Henrich Joseph, Steven J. Heine, and Norenzayan Ara. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.

[22] Yamuna Kachru. *Hindi*, volume 12. John Benjamins Publishing, 2006.

[23] Diptesh Kanojia, Kevin Patel, and Pushpak Bhattacharyya. Indian language wordnets and their linkages with princeton wordnet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[24] Pawan Lahoti, Namita Mittal, and Girdhari Singh. A survey on nlp resources, tools and techniques for marathi language processing. *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[25] Viet-Bac Le and Laurent Besacier. Automatic speech recognition for under-resourced languages: application to vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1471–1482, 2009.

[26] Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. Toward more meaningful resources for lower-resourced languages. *arXiv preprint arXiv:2202.12288*, 2022.

[27] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463, 2014.

[28] Mathieu Mangeot and Hong Thai Nguyen. Building lexical resources: towards programmable contributive platforms. In *2009 IEEE-RIVF International Conference on Computing and Communication Technologies*, pages 1–8. IEEE, 2009.

[29] Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[30] George A Miller. *WordNet: An electronic lexical database.* MIT press, 1998.

[31] Nandu Chandran Nair, Maria-chiara Giangregorio, and Fausto Giunchiglia. Is this enough?-evaluation of malayalam wordnet. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–108, 2021.

[32] Nandu Chandran Nair, Rajendran Sankara Velayuthan, and Khuyagbaatar Batsuren. Aligning the indowordnet with the princeton wordnet. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 9–16, 2019.

[33] Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*, 2002.

[34] Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, 2013.

[35] Ian Niles and Adam Pease. Mapping wordnet to the sumo ontology. In *Proceedings of the ieee international knowledge engineering conference*, pages 23–26, 2003.

[36] Rajeshwari Pandharipande. Minority matters: issues in minority languages in india. *International Journal on Multicultural Societies*, 4(2):213–234, 2002.

[37] S Rajendran and KP Soman. Malayalam wordnet. In *The WordNet in Indian Languages*, pages 119–145. Springer, 2017.

[38] Jaya Saraswati, Rajita Shukla, Ripple P Goyal, and Pushpak Bhattacharyya. Hindi to english wordnet linkage: Challenges and solutions. In *Proceedings of 3rd IndoWordNet Workshop, International Conference on Natural Language Processing 2010 (ICON 2010)*, 2010.

[39] Kevin P Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15, 2007.

[40] Archana Shukla and Varun Yadav. Spirituality and personality: Drawing parallels. *Journal of Indian*, page 111.

[41] Meghna Singh, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, and Pushpak Bhattacharyya. Mapping it differently: A solution to the linking challenges. In *Eighth Global Wordnet Conference*, 2016.

[42] Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. Automatic generation of english vocabulary tests. In *CSEDU (1)*, pages 77–87, 2015.

[43] Mark Turin. *Linguistic diversity and the preservation of endangered languages: A case study from Nepal.* International Centre for Integrated Mountain Development (ICIMOD), 2007.

[44] Titela Vîlceanu. Quality assurance in translation. a process-oriented approach. *Romanian Journal of English Studies*, 14(1):141–146, 2017.

[45] Piek Vossen. Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex. *international journal of Lexicography*, 17(2):161–173, 2004.

[46] PJTM Vossen. Eurowordnet. 1999.

[47] George Weber. Top languages. *The World' s*, 10, 2008.