Structural Analysis of Linking Urdu WordNet to PWN 2.1

Ayesha Zafar*, Afia Mahmood**, Sana Shams*, Sarmad Hussain*

* Center for Language Engineering, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, ** University of Education, Lahore
firstname.lastname@kics.edu.pk, afiamahmood@ue.edu.pk

Abstract

Multiple cross language WordNets such as EuroWordNet (EWN), Multi WordNet, Asian WordNet and Indo WordNet, have been developed that involve mapping Princeton WordNet (PWN) with the respective language WordNet [1,2,3,4,5]. Majority of these projects have employed the transfer-and-merge method developed during the construction of EuroWordNet for WordNet linkage. This paper discusses the process, challenges and results of linking Urdu WordNet, to the Princeton WordNet Version 2.1 from a linguistic and lexicographic perspective. Based on the synset alignment experience, cross language (Urdu – English) linkage issues have been highlighted followed by a contextual strategy for the resolution. Urdu language concepts that could not be aligned with the PWN 2.1 are also highlighted and discussed.

1 Introduction

WordNet is a lexical resource whose design is based on psycholinguistic theories of human memory on the one hand and the British school of structural/lexical semantics on the other [6]. Nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept [7]. There are semantic and lexical relations between lexical items which dominate their organization and exhibit their meaning. Moreover, these relations occur more often between words belonging to the same part of speech, thus nominal lexical items are networked with other nominal lexical items, verbal lexical items with verbal ones, etc. Furthermore, it is not composed of entries in the traditional lexicographical sense. WordNet assumes that synonyms grouped in synsets stand for concepts, and that most relations stick to concepts rather than to single lexical items [8].

Urdu WordNet is the first semantic dictionary of Urdu developed by Center for Language Engineering. It contains 5058 senses. All synsets have POS definition, unique synset ID, definition, synset and example. The example of an entry has been given in Figure 1.

Figure 1: Layout of Urdu WordNet

Increasing number of language specific WordNets has created interest in the linkage of WordNets to Princeton WordNet to enhance their usability. The linkage of synsets of one language to the other facilitates the development of bilingual dictionaries which can be used for machine translation and cross language information retrieval. It also alleviates the performance of word sense disambiguation tasks even in the absence of sense tagged corpora in a target language [3, 5, 9]. This paper reports the research challenges of aligning Urdu synsets with English synsets of PWN 2.1.

The paper is organized in the following sections. Section 2 reviews the current literature regarding various WordNet linkage projects and their reported accuracy statistics. Section 3 describes the approach of linking Urdu WordNet with PWN 2.1. Sections 4 presents in detail the challenges and solutions for linking Urdu concepts with English

1 http://cle.org.pk/clestore/urduwordnet.htm
2 Literature Review

Recently there have been multiple attempts to build WordNets for different languages and to link these WordNets to English WordNet. The process of linking involves the matching of a particular synset in one WordNet to a synset in another WordNet and requires high level of accuracy especially when the two languages belong to different cultures. In addition, conceptual gaps and the difference in the coarseness of the word senses are further challenges faced during alignment. As reported in [10], three types of difficulties were faced during the alignment of Romanian WordNet (RoWN) to PWN: (i) Difficulties caused by similar or intersecting synsets and non-differentiating or insufficiently distinguishing examples in PWN (ii) Difficulties caused by the structural differences in wordnet development, e.g., all word senses in PWN are equal, while Romanian wordnet has main and derived senses. Some idiomatic expressions are also missing in the Romanian wordnet (iii) Difficulties caused by the intrinsic differences between English and Romanian language i.e. at times English language meanings are missing in the Romanian language and vice versa.

Similarly, challenges were faced in the linkage of Hindi WordNet to PWN [11]. Hindi WordNet used a semi-automated system, WNSynsetMatcher tool [12], for linking the Hindi WordNet with the English WordNet. They describe that the main challenges faced were due to cultural difference in the concepts of kinship relations, musical instruments, grains, kitchen utensils, different tools and certain species of birds and animals. The solution proposed for alignment is using direct and hypernymy linkages.

The construction of Ancient Greek WordNet (AGWN) was automatic in which Greek-English digitized lexicons were used to extract Greek-English word pairs [13]. Later, the Greek word of the extracted pair was linked to every synset in the PWN. However, all the synsets of Greek were not available in the PWN. Thus, the AGWN contains 35,000 distinct lemmas with coverage of 28% of Greek lexicon, whereas the Greek lexicon contains 120,000 distinct lemmas. Bizzoni [13] state that English is polysemic in nature and the high polysynthetic nature of English and the relatively isolating character of the Greek contributed to major difficulties in the development of AGWN.

Thai WordNets have been constructed using the manual and semi-automated approach [14] [15]. This WordNet contains 21,344 senses. The major difficulties in the alignment of Thai WordNet to PWN were caused due to the conceptual gaps between Thai and English language. For example the meaning of retail store and store is opposite in Thai. Retail store denotes store and store denotes to retail store. Similarly, device, implement, tool, equipment etc. are mapped on only two words of Thai. Furthermore, one English word ‘doctor cannot be mapped on two genders.

Persian WordNet which is also aligned with PWN was created using the automatic approach. The approach used bilingual dictionary as well as Persian and English corpora to align the Persian and PWN synsets. Montazery et al [16] elaborate the method that their approach calculates a score for each candidate synset of a given Persian word and for each of its translations, it selects the synset with maximum score as a link to the Persian word. They report that this method brought more accuracy than the manual method. The accuracy of automatic approach has been reported as 82.6%.

Chinese [17] and Spanish [18] WordNets have been created using the automatic methods. Thai [15] and Hindi [11][12] WordNet have been developed using the semi-automatic approaches. Urdu WordNet [19] has been developed using the merge approach and later manual linkage of Urdu synsets to PWN 2.1 synsets. The following section presents the procedure of aligning Urdu WordNet with PWN 2.1 and consequently provides in detail the specific alignment challenges faced in the process.

3 Urdu WordNet to PWN 2.1 alignment methodology

5000 nouns, verbs, adjectives and adverbs were used to develop the Urdu WordNet (UWN) [19]. In the next stage, these 5000 words were reviewed and aligned to PWN 2.1. The following steps were followed during this process.

1. Firstly, the finalized Urdu Synset with its specific POS, relevant details of concept definition, example sentence and a unique ID was entered into the Urdu WordNet application. During Urdu synset finalization it was verified that all the senses of a specific synset were distinct (different from each other) and comprehensive (i.e. embody precise and adequate detail) for concept explanation.

2. Next, the verified Urdu senses were looked up in
the dictionaries for all the possible translations. Based on this lookup, at least three candidate words were to be selected for possible mapping.

3. Once the English candidate terms are generated, the complete POS category of its respective sense is carefully analysed. For example, Urdu senses depicting a state in the concept definition would be mapped to noun.state sense of the English word rather than noun.act or noun. artifact senses of the same word for consistency.

4. Once an English sense is finalized for mapping, its PWN sense ID is recorded against the particular Urdu sense. The following table shows the process.

Table 1: Urdu to English sense mapping

<table>
<thead>
<tr>
<th>Urdu Word</th>
<th>Urdu POS</th>
<th>Urdu Concept</th>
<th>Urdu Sentence</th>
<th>English Concept</th>
<th>English POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cristiano</td>
<td>N</td>
<td>جھاڑ گھاڑ</td>
<td>اس کا نے اپنی</td>
<td>Tranquility</td>
<td>noun.state</td>
</tr>
<tr>
<td>Ronaldo</td>
<td>N</td>
<td>خپڑہ کرنا</td>
<td>اس کا نے اپنی</td>
<td>repose</td>
<td>noun.state</td>
</tr>
</tbody>
</table>

5. Lastly, the selected candidate word is entered in the Urdu-English alignment utility. The utility displays all the senses of the selected word, and there the selected sense is selected to complete the mapping process.

4 Alignment challenges and proposed solutions

During the alignment of UWN to PWN 2.1 challenges faced were that of equivalence. These issues can be broadly categorized as syntactic, morphological and semantic differences. The following section discusses these alignment challenges and proposes solutions for alignment.

4.1 Morphological issue: Causative difference between Urdu and English

Urdu is morphologically richer than English as it has morphological devices such as inflection, that change verbs into their causative forms. Causitivization [20] is a process in which subject takes new arguments that changes the meaning of the verb. In Urdu, infixes like ل (lā-) and د (dā-) create verb causatives. Verbs in Urdu language are categorized into three forms which (i) Verb / ل (lā) / and (ii) Di-transitive Verb / د (dā) / and (iii) Di-transitive Verb / د (dā) / د (dā).

As shown in the table above, سوانا (s o:n a:/ sleep) is a root verb and its causative is سوانا (s o:n a:/ sleep) / preserve, to make someone sleep. In contrast, morphological causatives are not found in English. Therefore, during the WordNet linkage, the causative verbs in Urdu couldn't be mapped appropriately on English verbs.

As shown in the table above, سوانا (s o:n a:/ sleep) is a root verb and its causative is سوانا (s o:n a:/ sleep). Urdu root verb becomes passive whereas its causative remains active. In this case, causative maps directly on English word. For example, the causative بجھنا (bādana: to play) of the base verb بجھنا (bādana:)

<table>
<thead>
<tr>
<th>Urdu Word</th>
<th>Urdu Concept</th>
<th>Urdu Sentence</th>
<th>English Concept</th>
<th>English POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cristiano</td>
<td>جھاڑ گھاڑ</td>
<td>اس کا نے اپنی</td>
<td>Tranquility</td>
<td>noun.state</td>
</tr>
<tr>
<td>Ronaldo</td>
<td>خپڑہ کرنا</td>
<td>اس کا نے اپنی</td>
<td>repose</td>
<td>noun.state</td>
</tr>
</tbody>
</table>

Thus, سوانا (s o:n a:/ sleep) maps on sleep. However, no possible word for سوانا (s o:n a:/ sleep) could be found from PWN.

Similarly, پچکنا (pīṭkāna:/ squeeze) is a root verb, that changes to پچکنا (pīṭkāna:/ compressed) due to causitivization, and in the process it also changes its meaning. Furthermore, it was also observed that at times, Urdu root verb becomes passive whereas its causative remains active. In this case, causative maps directly on English word. For example, the causative بجھنا (bādana: to play) of the base verb بجھنا (bādana:)}
4.2 Syntactic issue: Complex predicates in Urdu causing POS mismatch in alignment

Another alignment challenge is faced due to complex predicates in Urdu as Urdu language employs different types of complex predicates to express its full range of verbal predication. [21] [22] Two types of complex predicates i.e. noun+verb and adj+verb were found common in the data which couldn’t be mapped.

In N+V and Adj+V complex predicates the noun and adjective contains the predicational content where as the verb, usually referred to the light verb [23]. For example, كرنانا (karna: / to disclose), and اثرناز بوتا (asaranːaːzhoː / to influence) are complex predicates in which nouns or adjectives require a verb to denote their complete meaning. They do not give complete meanings in isolation. In the examples given above بوتا (hoː na) and كرنانا (karna) are used to convey the complete meaning thus افشا افشا (afʃaː) ن威尔ا (N威尔ا) will always be used with كرنانا (karna) and اثرناز (asaranːaːz) Adj will always be used with بوتا (hoː na). This is presented in table 3 below.

Even though the complex predicates structurally comprise of two words, syntactically and semantically they behave like single constituents. Other examples of this issue are برواری (boraːbiri) ن威尔ا + كرنانا (karna) و.V.

The UWN to PWN alignment challenge arises when a noun in Urdu as always gives meaning of a verb. Therefore, it becomes confusing to map it with an English verb or English noun. As a solution such N/Adj+V constructions can be aligned with WordNet by adopting either list based approach or Rule based approach. However, complex predicates are considered highly productive with respect to their combinatorial possibilities. This means it is impossible to construct a static list of N/Adj+V combinations [23] [24]. In this scenario, it is useful to investigate the actual syntactic and semantic characteristics behind complex predicate formation [24]. Thus rule-based approach is recommended Using the rule based approach, heuristic are drawn from the semantic and syntactic features of the N/Adj + V constituents in a complex predicate. These generalizations are then used to predict the nature of these complex N/Adj+V constructions on the basis of the semantic features of the nouns or adjectives involved.

4.3 Semantic Issues

The following sub sections present detail of the semantic challenges faced in alignment of UWN to PWN

<table>
<thead>
<tr>
<th>Urdu Concept</th>
<th>Urdu Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>كرسن</td>
<td>he has revealed his secret to all</td>
</tr>
<tr>
<td>كرسن ولا</td>
<td>he has revealed his secret to all</td>
</tr>
<tr>
<td>كرسن ن威尔ا</td>
<td>he has revealed his secret to all</td>
</tr>
<tr>
<td>كرسن ولا</td>
<td>he has revealed his secret to all</td>
</tr>
</tbody>
</table>

4.3.1 Single Urdu concept for multiple PWN concepts:

During alignment, it was observed that some Urdu words in a particular sense could be mapped to multiple senses of a certain English word. For example, UWN Entry: <10028> <noun.time> (the time of person's life when they are a child)

Table 3: The Case of Complex Predicates

<table>
<thead>
<tr>
<th>Urdu Word</th>
<th>Urdu POS</th>
<th>Urdu Concept</th>
<th>Urdu Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>افشا</td>
<td>N</td>
<td>كرسن</td>
<td>he has revealed his secret to all</td>
</tr>
<tr>
<td>كرسن ولا</td>
<td>he has revealed his secret to all</td>
<td></td>
<td></td>
</tr>
<tr>
<td>كرسن ن威尔ا</td>
<td>he has revealed his secret to all</td>
<td></td>
<td></td>
</tr>
<tr>
<td>كرسن ولا</td>
<td>he has revealed his secret to all</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kam sinhna: ne: ki: karna, no: umari: is a noun in Urdu which can be accurately mapped on the following two different senses of the word childhood from PWN: {14948030} <noun.time> (the time of person's life when they are a child)
\{14235403\} <noun.state> (the state of a child between infancy and adolescence)

This is because بچہ (bətʃə) gives a generalized sense of childhood. Thus both noun.time sense and noun.state senses of the word childhood can be mapped. Similarly, UWN Entry: <100902/ka:n̪a/echidna/> is a noun in Urdu which can be accurately mapped on the following two senses of Echidna in PWN:

1. {01853520} <noun.animal> (a burrowing monotronte mammal covered with spines and having a long snout and claws for hunting ants and termites; native to New Guinea)
2. {01853149} <noun.animal> (a burrowing monotronte mammal covered with spines and having a long snout and claws for hunting ants and termites; native to Australia)

This alignment challenge can be handled through one to many mapping of concepts. The Urdu sense which composes multiple concepts of PWN in terms of their relations and general understanding can be aligned with all those senses of PWN.

4.3.2 Multiple Urdu concepts for single PWN concept

Another alignment challenge faced during UWN to PWN mapping was that multiple concepts of a particular Urdu word could be mapped on one word of English. For example, Urdu verb بچہ نا:ک (bətʃə:n̪a: /scared) has two senses in UWN;

بچہ نا:ک بنا جانور کا تھر کری ای بگز کر پھرکنا (bətʃə:n̪a:k d̪a:n̪a:k j̪a: bətʃə:n̪a:k ɡəːh hət̪n̪a: /animals scared and retreats)

and


Here, both the senses can be mapped on scared, {01762161} <verb.emotion> (cause fear in)

Thus as a solution it is proposed that both the Urdu concepts are aligned to a single PWN concept to resolve such semantic issues.

4.3.3 Difference in personal relationship

Urdu language organizes kinship terminologies in classificatory terms whereas English language uses descriptive terms for relationship. Family relation hierarchies are different in Urdu and English. This difference causes alignment challenge because the kinship terminologies in Urdu have a wider array of relationships that do not have corresponding senses in PWN. These are explained in the following three types of relationships:

- **Blood relations**

Urdu language carries different terms for blood relations, e.g., nephew in PWN is used as a son of your brother or sister whereas in UWN پہا:ناچ چا:یا (PWN چا:یا) means sister’s son and پہا:ناچ چا:یا (PWN چا:یا) is used for brother’s son.

Similarly, niece is a daughter of your brother or sister in English but پہا:ناچ چا:یا:یا (PWN چا:یا:یا) is sister’s daughter and پہا:ناچ چا:یا:یا (PWN چا:یا:یا) is brother’s daughter in Urdu. Moreover, a concept for brother’s wife, called پہا:ناچ چا:یا:یا (PWN چا:یا:یا) in Urdu and sister’s husband called پہا:ناچ چا:یا:یا (PWN چا:یا:یا) in Urdu is inexistent in the PWN 2.1. These differences represent lexical gaps in structuring of information in the case of blood relationships.

- **Relations with In-laws**

Urdu lexicalizes the distinction between the blood relations of husband and wife. However in English only two senses for these relations exist, {09731744} <noun.person> a brother by marriage and {10444395} <noun.person> the sister of your spouse whereas in Urdu، بھنگی (sə:l چا:ل) is used for wife’s brother and two terms are used for husbands’ brothers i.e. پہا:ناچ چا:یا:یا (PWN چا:یا:یا) elder brother of husband and پہا:ناچ چا:یا:یا (PWN چا:یا:یا) younger brother of husband. Also سالی (sə:l چا:ل) is used for wife’s sister and سالی (sə:l چا:ل) is used for husband’s sister.

- **Maternal and paternal relations**

There was another challenge for mapping maternal and paternal relationships. This is because English does not have specific concepts for relationships. For example پہا:ناچ چا:یا (PWN چا:یا) younger paternal uncle، پہا:ناچ چا:یا:یا (PWN چا:یا:یا) elder paternal uncle، پہا:ناچ چا:یا:یا (PWN چا:یا:یا) maternal uncle، په:ناچ چا:یا (PWN چا:یا) husband of mother’s sister and په:ناچ چا:یا (PWN چا:یا) husband of father’s sister all relations have only one corresponding English sense, uncle {10575646} <noun.person> -- (the brother of your father or mother; the husband of your aunt) in PWN. Similarly it is challenging to map other such relations like aunts, cousins, grandparents and grand-children where Urdu gives more than one sense for each of them based on gender, paternal and maternal side relations, separately.

This specific challenge of mapping personal relationships from Urdu language to English can be resolved by constructing hypernymy linkage. This means that in the absence of the equivalent English concept, the nearest term capturing the sense would be assumed as the hypernym of that concept and would be mapped to it. For example، مامون (mə:mən) and پہا:ناچ چا:یا (PWN چا:یا) would be mapped to the English synset of uncle.
4.3.4 Differences in representation of utensils

It was observed during mapping that certain kitchen utensils depict a category of words which is related to food, cooking and eating habits of the indigenous culture. For example, بِرْتُن (b r t n) mean kitchen ware, utensils made of clay, metal or glass; equipment for cooking and eating. In this case, بِرْتُن (b r t n) represent a composite sense of various utensils where as a sense capturing this concept in PWN could not be found. This is similar to the concept cutlery (a composite of spoons forks, etc.) for which we do not have a corresponding equal concept in Urdu. Similarly, ہُنگ (d : ŋ gu) is a culture specific sense that implies (i) کَذَا - کَذَا جَمِّدَ (lakrqi: ka: b r a: ʃəmə a: ) large wooden spoon, (ii) بِرْتُن جَسَ مِس شُورِیا وَغیرہ دستِخانہ بر جنگی بیس (b r t n d j s m s šūriya ɬəpər ɬəfər blən ʃət̪ xənə lı: hə: ) a bowl for curry, (iii) کمی وِژن بِرْتُن سے پیمانہ نکااگی کَذَا دار پِالیا کی شکل کُہوْنا ظرف (k ɬ i: b r t n se: paːni: nɪkəːnə: kaː dəndiː : ə r pj aː iː ɬ iː ʃəkəlkəː ʃət̪ xənə ʒərəf) a pot, which is used to extract water from any vessel. However, PWN only gives a general concept of utensil i.e. {04462854} <noun.artifact> an implement for practical use (especially in a household). Such issues can also be handled through direct linkage or hyponymy linkage. For example, the assumed hyponym of ہُنگ would be tableware (articles for use at the table (dishes and silverware and glassware)).

4.3.5 Differences in representation of fruits

There are many fruit names which are culture specific and are discretely lexicalized in Urdu. For example, گُری (kaːr iː ) unripe mango fruit is commonly used in Urdu. This issue can be handled by direct linkage. For example, گُری can be linked to English synset mango.

5 Un-mapped lexical and cultural senses

The different categories of alignment challenges discussed above can be resolved by adopting the proposed solution, however, some Urdu senses still remain unmapped. This is because of the inevitable linguistics, cultural, semantic differences of Urdu and English language. Few categories of these senses that remain unmapped are discussed below.

5.1 Cultural specific vegetables and utensil names

There are a few vegetables which cannot be mapped on any of the PWN senses as they only exist in Urdu, e.g. سَکُ سْ (s aː ə: g), بِنَہوں (biː : ʃ H uː a: ) remained unmapped due to the unavailability of proper concept in PWN. Similarly, there are certain utensils which only exit in Urdu, e.g. ہُہُرُ (b ə ʃ H uː ə: ) large drum of clay which is used to store grains, ہُنگ (d : 1) a medium size wooden spoon used for cooking.

5.2 Semantic orientation of borrowed words

Urdu has borrowed many words from English language. While mapping, it was revealed that the semantics of such English words when used in Urdu has changed and it does not give the same area of meaning as that of the originally borrowed foreign word. For example, یوہست (poːst/ any office or rank), is a borrowed word from English, but it could not be aligned to any of the PWN senses of the word ‘Post’.

Another example of different semantic orientation of borrowed words is اسَر (asər/ an officer) who has right to order. The Urdu concept of this word is not available in any of the PWN senses of the English word Officer being, {10216432} <noun.person>, someone who is appointed or elected to an office and who holds a position of trust. This semantic change refers to semantic shift or progression and involves changes in the usage of words where its literal sense radically differs from its original meaning. Moreover, such words couldn’t be mapped to a sense of a different English word.

5.3 Literal Concepts

There are many words in Urdu language based on stereotypes and culturally-inherited associations. Such metaphors do not hold true in all situations as are used as phrasal words. These also remained unmapped as no parallel senses exist in English. Table 4 illustrates few examples of these senses.

Table 4: Example of missing literal concepts

<table>
<thead>
<tr>
<th>Words</th>
<th>Concept</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
<tr>
<td>بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
<td>بھوْنا: بھوْنا:</td>
</tr>
<tr>
<td>bhuːn:</td>
<td>bhuːn: bhuːn:</td>
<td>bhuːn: bhuːn:</td>
</tr>
</tbody>
</table>
5.4 Un-categorized conceptual gaps in Urdu and English

There are many concepts in Urdu which remain unmapped due to unavailability of corresponding concepts in PWN. These concepts are of varied nature thus, un-categorized and tabulated below.

<table>
<thead>
<tr>
<th>Table 5: Un-categorized conceptual gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Words</strong></td>
</tr>
<tr>
<td>parva:z</td>
</tr>
<tr>
<td>baza:ri:</td>
</tr>
<tr>
<td>a:na</td>
</tr>
</tbody>
</table>

In the table above, پردر:ز (parva:z/ flight) is an Urdu concept depicting پردر:ز کی آران (parma:kı: or:a:n / flight of birds), which could potentially be mapped to flight. However, it was observed that flight gives a generic concept of flying, whereas Urdu WordNet provides a specific concept for flight of birds which is not available in PWN. Similar patterns are observed in other Urdu words as well.

6 Alignment Results

The current status of English-Urdu aligned senses have been given in table 6 below. During the alignment process total 3526 Urdu senses from UWN have been reviewed out of which 1829 Urdu senses were aligned to PWN 2.0. This is shown in table below.

<table>
<thead>
<tr>
<th>Table 6: WordNet data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number of reviewed senses</strong></td>
</tr>
<tr>
<td>Total number of UWN senses</td>
</tr>
<tr>
<td>Total number of senses aligned to PWN 2.0</td>
</tr>
<tr>
<td>Total number of unmapped senses</td>
</tr>
</tbody>
</table>

Within the total 1829 senses aligned, the following table provided the total count of nouns, adjectives and verbs.

<table>
<thead>
<tr>
<th>Table 7: Count of Aligned Senses as per Parts of Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total count of Nouns, Adjective and Verbs from UWN</strong></td>
</tr>
<tr>
<td>Total number of Nouns</td>
</tr>
<tr>
<td>Total number of Adjectives</td>
</tr>
<tr>
<td>Total number of Verbs</td>
</tr>
</tbody>
</table>

1403 Urdu sense remained unmapped due to cultural, religious, semantic and linguistic differences. The percentage of unmapped senses is 39.79 % which is higher in number. The issues of unmapped senses have already been discussed. On the basis of proposed suggestion, these unmapped senses will be further reviewed and attempted to be aligned to PWN 2.1 through continued research. The work accomplished to data is available at CLE’s² website.

7 Conclusion

This paper reports the UWN to PWN mapping methodology, issues and challenges while aligning Urdu WordNet to PWN. It was observed that morphological, syntactical, semantic and cultural issues were a hindrance in accomplishing Urdu to English mapping. However, possible solutions are suggested to resolve these issues. Further research needs to be conducted in hypernym relationship development and Urdu VerbNet development in order to resolve the alignment challenges for effective alignment.

8 Acknowledgements

This work has been conducted through the project, Essential Linguistic Resources project³, supported through a research grant from DAAD, Germany.

² http://www.cle.org.pk/clestore/urduwordnet.htm
³ http://www.cle.org.pk/clestore/
9 References


