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 Euro WordNet (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 Euro WordNet 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[1] 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.

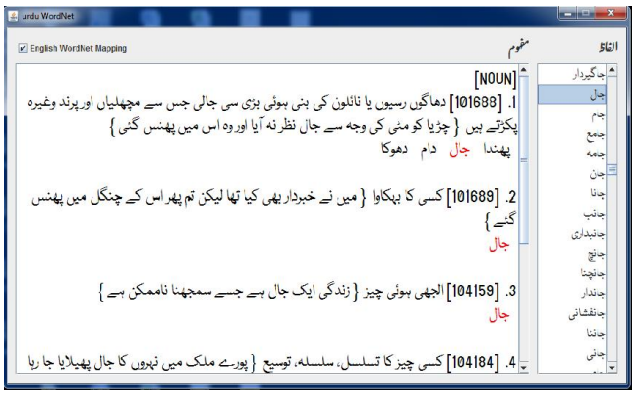

Figure 1: Layout of Urdu WordNet

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

synsets. Section 5 documents concept categories that remain un-linked. Alignment results are discussed in section 6. Finally, Section 7 concludes the paper by reporting the future work required in this direction.

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.

Similar 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 wordfor 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

Urdu Word	امن/	امن	امن
Urdu POS	N	N	N
Urdu Concept	چین اور اطمینان سے بھرپور ماحول یا جگہ (پُر کے ساتھ)	جنگ کی ضد	سکون کی حالت
Urdu Sentence	بہت جلد اسے دشمنوں سے نجات کی صورت میں ایک پُر امن جگہ مل گئی	اس بار جنگ وجدل سے کام لینے بجائے ہم امن کا پیغام لے کر جائیں گے	امیر شہر کے آتے ہی لوگ امن سے رہنے لگے
Candidate Terms	peace, repose, reconciliation	peace, harmony, accord	Tranquility, calm, serenity
Selected Eng Word	Peace	peace	tranquility
Eng Concept	the absence of mental stress or anxiety	the state prevailing during the absence of war	an untroubled state; free from disturbance
POS in PWN	noun.feeling	noun.state	noun.state
Eng Sense ID	07413685	13784195	13783084

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 لا (lā-) and وا (vā-) create verb causatives. Verbs in Urdu language are categorized into three forms which (i) Verb/ لازم / la:z ɪ m (ii) Transitive Verb/متعدی / mu ʈ ə ɖ ɪ / and (iii) Di-transitive Verb /متعدی المتعدی / mu ʈ ə ɖ ɪ ə l mu ʈ ə ɖ. In most of the cases, (i) لازم represents the root verb, while (ii) متعدی and (iii)متعدی المتعدی represents its causatives. It is shown in the following table 2.

Table 2: Examples of Urdu root verbs and their causatives

متعدی المتعدی (di-transitive verb)	متعدی (transitive verb)	لازم (root verb)
سلوانا sʊlvɑ:nɑ:	سلانا sʊlɑ:nɑ:	سونا s ɔ:n ɑ:
بجوانا bədʒvɑ:nɑ	بجانا bədʒɑ:nɑ	بجنا bədʒnɑ:
پچکوانا pɪʧəvɑ:knɑ:	پچکانا pɪʧəknɑ:	پچکنا pɪʧəknɑ:

As shown in the table above, سونا (s ɔ:n ɑ:/ sleep) is a root verb and its causative is سلانا (sʊlɑ:nɑ:/ 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.

UWN Entry: <100795><سونا/ s ɔ:n ɑ: /sleep><N >< نیند آجانا/ n i:nd̪ ɑ:dʒɑ:nɑ:/ be asleep>< بچہ سونا چاہتا ہے/bəʧʧɑ: s ɔ:n ɑ:ʧɑ:hʈɑ: h æ: / the baby wants to sleep>
PWN Entry: {00014762} <verb.body> (be asleep)

Thus, سونا maps on sleep. However, no possible word for سلانا could be found from PWN.

Similarly,پچکنا (pɪʧəknɑ:/ squeeze) is a root verb, that changes to پچکانا (pɪʧkɑ:nɑ:/ 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 بجانا (bədʒɑ:nɑ/ to play) of the base verb بجنا (bədʒnɑ:/

automatic play) is mapped on Play <01710937>, where as the base word automatic play (bəʤnɑ:/بجنا) remains unmapped. Similar phenomenon can be observed in other Urdu verbs like, پچکنا (pɪʧəknɑ/get squeezed) and بٹنا (bətnɑ:/ get distributed)

These issues can be handled through VerbNet. VerbNet associates the semantics of a verb with its syntactic frames, and combines traditional lexical semantic information such as thematic roles and semantic predicates, with syntactic frames and selectional restrictions. Therefore, such causative verbs can be clustered in semantically coherent classes. Verb lexicon which is based on VerbNet can be linked to WordNet.

4.2 Syntactic issue: Complex predicates in Urdu causing POS mis-match 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,کرناافشا (əfʃɑ: kərnɑ:/ to disclose), and اثرانداز ہونا (əsərənɖɑ:zho: na:/ 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 ہونا (ho: na) and کرنا (kərnɑ) are used to convey the complete meaning thus افشا (əfʃɑ) N will always be used with کرنا/V and اثرانداز (əsərənɖɑ:z) Adj will always be used with ہونا/V. This is presented in table 3 below.

Table 3: The Case of Complex Predicates

Urdu Word	Urdu POS	Urdu Concept	Urdu Example
افشا əfʃɑ: reveal	N	کسی چیز کو ظاہر یا عیاں کرنے کا عمل k ɪsi: ʧi:zko: zɑ:hır j ɑ: əj ɑ̃: kərne: kɑ: əməl the act of displaying anything	اس نے اپنا راز سب پر افشا کر دیا ʊs ne: əpn ɑ: rɑ:zsəbpərəfʃɑ: kərɖɪj ɑ: he has revealed his secret to all
اثرانداز əsərən ɖɑ:z affect	Adj	اثرڈالنے والی /والا əsərdɑ:lne: vɑ:lɑ: putting an affect	ہمارے ملک کی آب و ہوا گرم ہے جلدی اثرانداز ہوتی ہے həmɑ:re: mʊlkki: ɑ:bo:h əvɑ: gərəmhæ: ʤəlɖi: əsərənɖɑ:z ho: ʈi: hæ: our country's climate is hot, it affects quickly

Even though the complex predicates structurally comprise of two words, syntactically and semantically they behave like single constituents. Other examples of this issue are برابری(bərɑ:bəri) N + کرنا (kərnɑ) V.

The UWN to PWN alignment challenge arises when افشا a noun in Urdu as always gives meaning of a verb. Therefore, it becomes confusing to map it with a 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

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: <100281><بچپن/ bəʧpən><کم سن ہونے کی کیفیت، نوعمری kəmsɪnhɔ:ne: ki: kæ:fɪjə ʈ , no: ʊməri:> 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 بچپن (bəʧpən) 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 ><کانٹا /kɑ:ntɑ:/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 monotreme 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 monotreme 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 بدکنا (bɪɖəknɑ:/ scared) has two senses in UWN;
<101339> پیچھے ہٹنا جانور کا ڈر کر یا بِگڑ کر بھڑکنا، ʤ ɑ:nvərkɑ:dərkər jɑ: bɪgəɽkər pi:ʧʰe hətnɑ:/ animals scared and retreats
and <101340> آدمی کا یکایک کسی سے ڈر کر بدگمان ہونا، الگ ہو جانا/ɑ:ɖmi: kɑ: kɪsi: se: dərkərbəɖgʊmɑ:n ho:nɑələg ho:ʤ ɑ:nɑ:/suddenly a man gets scared to be skeptical
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 بھانجا (bʰɑ:n ʤ ɑ:) means sister's son and بھتیجا (bʰəʈi: ʤ ɑ:)is used for brother's son.
Similarly, niece is a daughter of your brother or sister in English but بھانجی (bʰɑ:nʤi:)is sister's daughter and بھتیجی (bʰəʈi: ʤi) is brother's daughter in Urdu. Moreover, a concept for brother's wife, called بھابھی (bʰɑ: bʰi:) in Urdu and sister's husband called بہنوئی(bʰɛno:i:) 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, سالا (sɑ:l ɑ:) is used for wife's brother and two terms are used for husbands' brothers i.e. جیٹھ(ʤe:tʰ) elder brother of husband and دیور (ɖe:v ə r) younger brother of husband. Also سالی (sɑ:li:) is used for wife's sister and نند (n ə n ɖ) 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 چچا (ʧ ə ʧ ɑ) younger paternal uncle, تایا (ʈ ɑ:j ɑ:) elder paternal uncle, ماموں (mɑ:m ũ:) maternal uncle, خالو (x ɑ:lu:) husband of mother's sister and پھوپھا (p^h ʊp^h ɑ:) 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 hypernymy of that concept and would be mapped to it. For example, چچا، تایا and ماموں 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 depicts a category of words which is related to food, cooking and eating habits of the indigenous culture. For example, برتن (b ə r ʈ ə n) mean kitchen ware, utensils made of clay, metal or glass; equipment for cooking and eating. In this case, برتن (b ə r ʈ ə 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 ڈونگا (d ɔ: ŋ gɑ) is a culture specific sense that implies (i) لکڑی کا بڑا چمچہ، کونڈا (ləkɽi: kɑ: bəɽ ɑ: ʧəmə ɑ:) large wooden spoon, (ii) برتن جس میں شوربا وغیرہ دسترخوان پر چنتے ہیں (bərʈən ʤ ɪs m ẽ: ʃo:rbɑ: və ɣ æ:rɑ: ʤəsʈərxɑ:npər ʧ ʊnʈe: hæ̃:) a bowl for curry, (iii) کسی بڑے برتن سے پانی نکالنے کا ڈنڈی دار پیالے کی شکل کا چھوٹا ظرف (k ɪsi: bəɽe: bərʈən se: pɑ:ni: nɪkɑ:lne: kɑ: dəndi: ʤ ɑ:r pɪj ɑ: le: ki: ʃəkəlkɑ: ʧo:t ɑ: zərf) 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 hypernymy linkage. For example, the assumed hypernymy of ڈونگا 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, کیری (kæ:ri:) unripe mango fruit is commonly used in Urdu. This issue can be handled by direct linkage. For example, کیریcan 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. ساگ (s ɑ: g), بتھوا (bi: ʈ H u: ə) , remained unmapped due to the unavailability of proper concept in PWN. Similarly, there are certain utensils which only exit in Urdu, e.g. بھڑولا (b^h ə ɽ ɔ:l ɑ:) large drum of clay which is used to store grains, ڈوئی (d ɔ: I) 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 پوسٹ (po:st/ 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 افسر (əfsər/ 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

Words	Concept	Example
بھوننا **bhu:nənɑ:**	بندوق سے گولیاں مارکر قتل کر دینا bənʤu:qse: go:lɪj ɑ̃: mɑ:rke: qəʈəlkərʤe:n ɑ:	فوج نے ایک ہی حملہ کیا اور دشمن کے کئی سپاہی گولیوں سے بھون دیے f ɔ: ʤ ne: ɪk hi: həmlɑ: kɪj ɑ: ɔ:r ʤʊʃmənko: bhu:nkərrəkʤɪj ɑ:
پھڑکنا **phʊɽəknɑ:**	غیر معمولی حرکت یا جنبش ɣæ:r m ɑ:mu:li: hərkə ʈ jɑ: ʤʊmbɪʃ	صبح سے ہی اس کی آنکھ پھڑک رہی تھی sʊbɑ:h se: hi: ʊski: ɑ:ŋkh phʊɽəkrəhi: ʈhi:

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 5: Un-categorized conceptual gaps

Words	Concept	Example
پرواز pərvɑ:z	پرندوں کی اڑان pərɪnd̪õ:ki: ʊɽɑ:n	مجھے وہ دن اچھی طرح یاد ہے جب ہمارے کبوتر نے پہلی پرواز کے لیے پر کھولے mʊd̪ʒhe: vo: d̪ɪnəʈʃʈʃî: ʈ̪ərhɑ: jɑ:d̪ h æ d̪ʒəbhəmɑ:re: kəbu:ʈ̪ər ne: pərvɑ:zke: lɪj e: pərkh o:le:
بازاری **bɑ:zɑ:ri:**	عامیانہ یا سوفیانہ، مبتذل، خواص کی نظر میں تہذیب سے گرا ہوا ɑ:mjɑ:nɑ jɑ: su:fjɑ:nɑ: , mʊbʈ̪əzıl x ɑ:s ki: nəzər m ẽ: ʈ̪ɛhzi:bse: gır ɑ: hʊvɑ:	بازاری گفتگو سے پرہیز کرو bɑ:zɑ:rigʊfʈ̪əgu: se: pərhe:zkəro:
آنہ ɑ:nɑ:	روپے کا سولہواں حصہ، جو قیمت میں ایک روپے کے سولہویں حصے کے برابر ہوتا ہے **r**ʊpe: k ɑ: so:lhvɑ̃: hissɑ: dʒo: qi:məʈ̪m ẽ: e:k **r**ʊpe:ke: so:lhvẽ: hisse: ke: bərɑ:bər ho:ʈ̪ɑ: hæ:	ہمارے دادا کے زمانے میں دو آنے کی روٹی ہوا کرتی تھی həmɑre: d̪ɑ: d̪ɑ: ke: zəm ɑ:ne: m ẽd̪o: ɑ:ne: ki: roti: huvɑ: kərʈ̪i: ʈ̪ʰi:

In the table above, پرواز (pərvɑ:z/ flight) is an Urdu concept depicting پرندوں کی اڑان (pərɪnd̪õ:ki: ʊɽɑ: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 6: WordNet data

Total number of reviewed senses	
Total number of UWN senses	3526
Total number of senses aligned to PWN 2.0	1829
Total number of unmapped senses	1403

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

Table 7: Count of Aligned Senses as per Parts of Speech

Total count of Nouns, Adjective and Verbs from UWN	
Total number of Nouns	1002
Total number of Adjectives	872
Total number of Verbs	249

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's[2] 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 project[3], supported through a research grant from DAAD, Germany.

[2] http://www.cle.org.pk/clestore/urduwordnet.htm
[3] http://www.cle.org.pk/eulr/

9 References

[1] P. Vossen, "EuroWordNet: a Multilingual Database for Information Retrieval", workshop on Cross-language Information, Zurich.

[2] E. Pianta, L. Bentivogli, C. Girardi, "MultiWordNet: Developing an Aligned Multilingual Database", In Proceedings of the *First International Conference on Global WordNet*, Mysore, India, 2002.

[3] V. Sornlertlamvanich, "Review on Development of Asian WordNet" Japio year book, 2009.

[4] J. Ramanand, A. Ukey, B. Singh, P. Bhattacharyya, "Mapping and Structural Analysis of Multi-lingual Wordnets", In Proceedings of *IEEE,* Bombay, 2007.

[5] G. Miller, Beckwidth, C. Fellbaum, D. Gross, "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography,* Vol 3, No.4 (1990), pages 235-244.

[6] C. Fellbaum , "*WordNet: An Electronic Lexical Database."* MIT Press, 1998.

[7] C. Fellbaum, M. Palmer, L. Delfs, S. Wolf, "Manual and Automatic Semantic Annotation with WordNet", In Proceedings of *NAACL,* Pittsburgh, 2001.

[8] P. Vossen , "EuroWordNet: A Multilingual Database with Lexical Semantic Networks." Kluwer Academic Publishers, Dordrecht, 1998.

[9] J. Daude, L. Padro, G. Rigau, "Mapping Wordnets Using Structural Information." 38th Annual Meeting of the Association for Computational Linguistics, 2000.

[10] D. Cristea, C. Mihiala, C. Forascu, D. Trandabat, M. Husarciuc, G, Haja, O, Postolache, "Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets", *Romanian Journal of Information Science and Technology. Volume 7,* 2004, (June, 27, 2014), pages 125-145

[11] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoor, A. Famian, S. Bagherbeigi, S. Assi, "Semi Automatic Development of Farsnet; the Persian WordNet", *In Proceedings of 5th Global WordNet Conference*, Mumbai, India, 2010.

[12] J. Saraswat, S. Ripple, P. Goyal, P. Bhattacharyya, "Hindi to English Wordnet Linkage: Challenges and Solutions".

[13] Y. Bizzoni, F. Boschetti, R. Del Gratta, H. Diakoff, M. Monachini, G. Crane, "The Making of Ancient Greek WordNet", In Proceedings of *Language Resources and Evaluation Conference,* Iceland, 2014.

[14] D. Leenoi, T. Supnithi, W. Aroonmanakun, "Building a Gold Standard for Thai WordNet", In Proceeding of *The International Conference on Asian Language Processing,* Thailand, 2008. Available at:

[15] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich, H. Isahara, "Thai WordNet Construction", In Proceedings of *the 7th Workshop on Asian Language Resources*, Association for Computational Linguistics, 2009.

[16] M. Mortaza, H. Faili, "Automatic Persian WordNet Construction." In Proceedings of *the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, Beijing, China, 2010.

[17] R. Xu, Z. Gao, Y. Pan, Y. Qu, Z. Huang, "An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet", In proceedings of *The Semantic Web*, Heidelberg, 2008.

[18] J. Atserias, L. Villarejo, G. Rigau, "Spanish WordNet 1.6: Porting the Spanish Wordnet Across Princeton Versions", In Proceedings of *Language Resources and Evaluation Conference,* Portugal, 2004.

[19] A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using the Merge Approach ", in the *Proceedings of Conference on Language and Technology* 2012 (CLT12), Lahore, Pakistan.

[20] S. M. J. Rizvi, *Development of Algorithms and Computational Grammar for Urdu (Ch 4: Urdu Verbs Characteristics and Morphology),* PHD thesis, Department of Computer & Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad. 2007.

[21] T. Mohanan, *Argument Structure in Hindi,* CSLI Publications, 1994.

[22] M. Butt, *The Structure of Complex Predicates,* PHD thesis, Department of Linguistics, Stanford University, 1995.

[23] M. Butt, T. Ahmed, "Discovering Semantic Classes for Urdu N-V Complex Predicates", in proc. *International Conference on Computational Semantics,* UK, 2011.

[24] M. Butt, T. Bogel, A. Haulti, S. Sulger, T. Ahmed, "Identifying Urdu Complex Predication via Bigram Extraction", in proc. *COLING 2012*, India, 2012.