Citation: A. Agostini, T. Usmanov, U. Khamdamov, N. Abdurakhmonova, M. Mamasaidov, "UZWORDNET: A Lexical-Semantic Database for the Uzbek Language. In S. Bosch, C. Fellbaum, M. Griesel, A. Rademaker and P. Vossen, editors, *Proceedings of the 11th International Global Wordnet Conference* (GWC-2021), pp. 8–19, Potchefstroom, South Africa, 2021.

UZWORDNET: A Lexical-Semantic Database for the Uzbek Language

Alessandro Agostini[1,2], Timur Usmanov[1],
Ulugbek Khamdamov[1], Nilufar Abdurakhmonova[3], Mukhammadsaid Mamasaidov[1]
[1]LDKR Group*, Inha University in Tashkent, Uzbekistan
[2]ISGS, Inha University, Incheon, South Korea
[3]Tashkent State University of Uzbek Language and Literature, Tashkent, Uzbekistan
a.agostini@inha.uz
{t.usmanov,u.khamdamov,m.mamasaidov}@student.inha.uz
abdurahmonova.1987@mail.ru

Abstract

The results reported in this paper aim to increase the presence of the Uzbek language in the Internet and its usability within IT applications. We describe the initial development of a "word-net" for the Uzbek language compatible to Princeton WordNet. We called it UZWORDNET. In the current version, UZWORDNET contains 28140 synsets, 64389 sense and 20683 words; its estimated accuracy is 75.98%. To the best of our knowledge, it is the largest wordnet for Uzbek existing to date, and the second wordnet developed overall.

1 Introduction

By living in the world, we—human 'agents'—and machines as well do not just make meanings up from language independently of the world. This is the *language problem* (Wittgenstein, 1953; Steels, 1997; Steels et al., 2002), and it is crucial for IT applications worldwide (Knight, 2016).

Unfortunately, computer scientists and engineers are still learning how to efficiently solve the language problem in their theories or applications, and understand how language-based technologies called "universal language models" work. They are often surprised by the mistakes that new AI tools are making.[1] In short, new technologies proliferate, and language-based biases appear increasingly almost anywhere in applications.

A problem of current technologies is that if a language is endangered, it is possible it will never have a life within them—and in the Internet online. Far from infinite, usable technology seems only as big as the language(s) we speak as users.

Language is just as important for building human connections online as it is offline: it forms the basis of how users identify with each other, the lines on which exclusion and inclusion are often drawn, and the boundaries within which communities grow around common interests.

As a consequence, the relationship between language diversity and the Internet is a growing area of policy interest and academic study.[2] The story emerging is one where language profoundly affects our experience of the Internet. It is a matter of fact, for instance, that Google searching for "restaurants" in English may bring us back 10+ times the results of doing so in another language.

For "another language", we focus on the Northern Uzbek language, a Turkic language officially recognized as the state language of the Republic of Uzbekistan. In particular, in this paper we advance and discuss initial results on the ongoing development of UZWORDNET (UZW in short), a prototypical version of a wordnet for the Uzbek language compatible to Princeton WordNet (Miller, 1995; Fellbaum, 1998).[3] Our long-term objective is to motivate, support and increase the study of computational aspects of Uzbek and, more generally, the usability of Uzbek within IT applications and the Internet. As a consequence, UZWORDNET is added to the Wordnets in the world[4] and provided open source under a license and format compatible with the Open Multilingual Wordnet (Bond and Paik, 2012; Bond and Foster, 2013).[5]

This paper is structured as follows. Below are some elements of the Uzbek language, followed by a brief excursus on word-nets (Section 3). In

* The acronym "LDKR" means Language, Data, Knowledge, and Reasoning. The LDKR Group aims to discovering (learning), modeling, reducing and computing the "semantic gap" between users and the Universe of Language(s), Data, Information and Knowledge their ICT systems are based on.

[1]For instance, see https://medium.com/@robert.munro/bias-in-ai-3ea569f79d6a (accessed 30 Nov 2019).

[2]For instance, see http://labs.theguardian.com/digital-language-divide/ (accessed 17 Oct 2019).

[3]https://wordnet.princeton.edu/.

[4]globalwordnet.org/resources/wordnets-in-the-world/.

[5]http://compling.hss.ntu.edu.sg/omw/.

Section 4, we focus on the few previous attempts towards the construction of a wordnet for Uzbek. In Section 5 we advance and discuss the work that produced UzWordnet. We validate and analyse the results in Section 7 and 8. We conclude with a summary and future work (Section 9).

2 Elements of Uzbek Language

Unless otherwise stated, in this paper by "Uzbek language" (native: *O'zbek tili*) we refer to the Northern Uzbek language. In fact, there is another Uzbek language—the Southern Uzbek—statutory language of provincial identity in Afganistan, spoken by about 6.5 million people wordwide (Eberhard et al., 2020).

Figure 1: Spread of Uzbek languages.

The (Northern) Uzbek language is a statutory national language in Uzbekistan.[6] It is a Turkic language and spoken by approximately 26.8 million people around the world (Ethnologue, 2020a), remarkably by a large group of ethnic Uzbeks residing abroad in Afghanistan, Kyrgyzstan, Kazakhstan, Turkmenistan, Tajikistan, Russia, Turkey, and Xinjiang (China), making it the second-most widely spoken Turkic language after Turkish (Ethnologue, 2020b). Figure 1 provides the rough geographical distribution of the Northern (majority) and Southern (minority) Uzbek languages.

[6]In spite of its status (1995, Official Language Law, amended, 3561-XI, Art.1), the Uzbek languange has been experimenting a number of issues for the disclosure of its full potentialities; see for instance cabar.asia/en/uzbekistan-why-uzbek-language-has-not-become-a-language-of-politics-and-science (accessed 12 Oct 2020).

The Uzbek languages are a descendant of Chagatai language, also known as the old-version of Uzbek. As a primary language of the Timurid dynasty, Chagatai represented the eclectic mixture of Turkic, Persian (or Farsi), and Arabic. After its extinction by the 19th century, its successor language lost its vowel-harmonization due to influence of Soviet standardization process (Hirsch, 2005) and became the standard (Northern) Uzbek we consider in this paper. Both languages belong to the Eastern subgroup of Turkic family, also known as the Karluk branch, along with the Uyghur language. In Figure 2, five most-spoken Turkic languages and their branches are depicted.

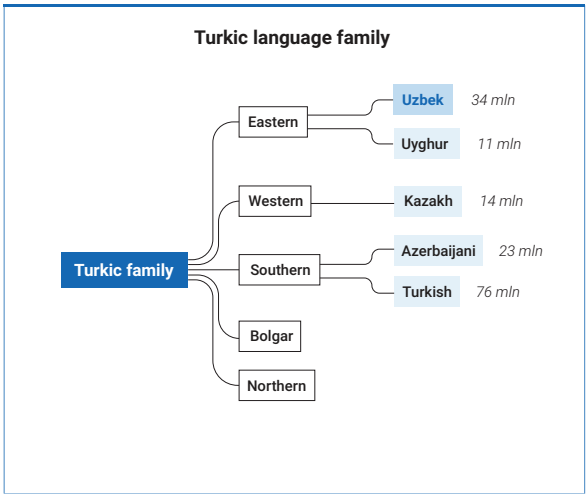

Figure 2: Widely spoken Turkic languages.

2.1 Dialects

The (Northern) Uzbek has three dialects: Karluk (or Karluk-Chigil-Uyghur), Kipchak, and Oghuz (Abdurokhmonov and Darvishev, 2011) (Figure 3). *Karluk* is a group of subdialects with a total number of 22-23 million speakers. It is divided into three main groups: Ferghana (covering almost the whole Ferghana Valley), Tashkent (the city and its region) and Qarshi, Samarkand and Bukhara groups. Karluk dialect became the standard form of Uzbek. *Kipchak* is a quite dispersed dialect. The total number of speakers is not yet calculated; that is to say, it accounts for the minority of speakers. Since the Karluk dialect is the standard on all levels of government and universities, the popularity of Kipchak is slowly declining. *Oguz* is spoken by approximately 2 million speakers, and it is widely spread in the Khorezm region, the Republic of Karakalpakstan, and the western part of the Bukhara region (To'ychiboev and Khasanov,

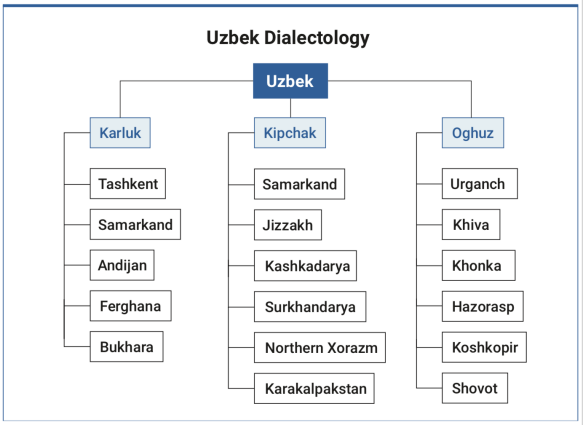

Figure 3: Uzbek dialects and their classification.

2004). Figure 4 is a roughly estimated visualization of Uzbek dialects spoken in Uzbekistan. Owing to the fact that Karluk and Kipchak dialects are dispersed throughout the country, each province is given the color of dialect of majority.

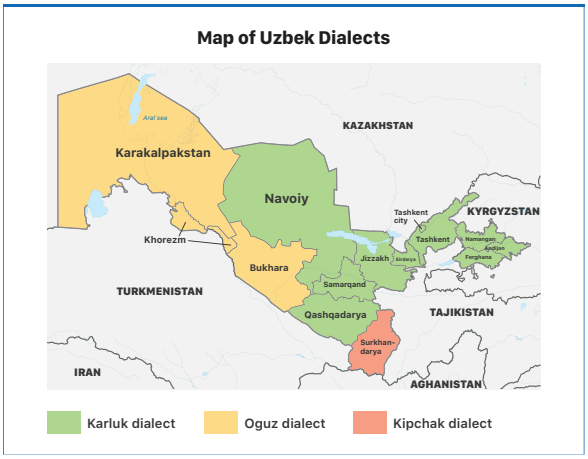

Figure 4: Map of (Northern) Uzbek dialects.

3 Word-Nets

In mid 1980s, several linguists and psychologists at Princeton Univesity started to model and develop a lexical-semantic database now referred to as Princeton WordNet (Miller, 1995; Fellbaum, 1998, PWN in short). The basic idea behind PWN has been to provide an aid in searching dictionaries conceptually, rather than merely alphabetically.

PWN divides the lexicon into four categories: nouns, verbs, adjectives, and adverbs. They are organized into synonym sets, each representing one underlying lexical concept (Miller et al., 1990).

PWN is based on synonyms and hyponyms of nouns and verbs, as well as antonyms of adjectives. In addition, it includes troponyms for verbs and hyponyms for nouns.

Princeton WordNet and a vast majority of wordnets (see for instance (Bond and Paik, 2012; Neale, 2018) for surveys and (Bond and Foster, 2013; Vossen, 1998) for extensions to open multilingual wordnet) were formed by expanding the semantic structure of PWN according to the *extend model*[7] (Vossen, 1998; Bond and Paik, 2012, p.67), which assumes that lemmas of the new language are created by translating English synsets of PWN. There is also possibility for creating semantic network by directly adding words and their definitions for the language under consideration. However, few wordnets have been created by using this method (*merge model*), due to the high cost of human expertise.

4 A Wordnet for Uzbek Language

Computational linguistics appeared as a field of research in Uzbekistan since the late 2000s; see for instance (Pulatov, 2011; Rakhimov, 2011; Abdurakhmonova, 2020, pp. 17-19). Since then, there have been few attempts to resolve lexical ambiguity in Uzbek by creating a sematic network, and none of them produced a word-net as we generally mean the term today after the pioneristic work at Princeton, see Section 3.

The very first wordnet for Uzbek—and, to the best of our knowledge, the only existing up to the present work on UzWordnet—is due to (Bond and Foster, 2013) as part of the Extended Open Multilingual WordNet project. The resulting wordnet (code **uzb**, accessible online[8]) is minimal in terms of available synsets and coverage of "core concepts" (Boyd-Graber et al., 2006):

Synsets	Words	Senses	Core concepts
889	1,115	1,157	8%

It also seems not clear what of two Uzbek languages the wordnet was built for. Moreover, we could not find a *specific* report on it (apart from the aforementioned numbers) and the estimated *specific* accuracy.[9]

In (Matlatipov et al., 2018), the authors focused on modeling a wordnet-like thesaurus for Uzbek, and tried to come up with a way to create rules for converting paper-based dictionary thesauruses

[7] Sometimes informally referred to as *expansion method*.
[8] http://compling.hss.ntu.edu.sg/omw/summx.html.
[9] The estimated accuracy is claimed to be 94% *over the 150+ languages* considered in the project.

into e-version using PROLOG. To develop a formal model of thesaurus, they built a dictionary's meta-language and defined its systematic properties. As a result, they obtained a *model*, not a wordnet as a computer system. (Abdurakhmonova and Khaydarov, 2019) surveys the main features of PWN towards its translation to Uzbek.

The list of works that explicitely target Uzbek (Northern or Southern) for the purpose of building a wordnet ends here. However, there have been numerous projects for other Turkic languages, for instance the development of a Turkish wordnet (Bilgin et al., 2004; Çetinoğlu et al., 2018). The project, started at Sabanci University of Istanbul as part of the BalkaNet project, uses a combination of the expansion and merge approaches. Another wordnet for Turkish is KENET (Ehsani et al., 2018). KENET is not based on PWN and is the most comprehensive wordnet for Turkish built from scratch using a bottom-up method. The wordnet was created by using the Contemporary Dictionary of Turkish (CDT) as lexical resource.

5 Our Approach

In this section we describe our approach to the construction of UZWORDNET. We divide it into three parts. First, the choice and pre-processing of the lexical resource. Second, the automatic construction of the PWN-like structure for (Northern) Uzbek, i.e., of UZWORDNET. Third, the expert human validation of the automatic construction.

5.1 Lexical resource

The lexical resource we used is the English-Uzbek Dictionary (*Inglizcha-O'zbekcha Lug'at*) by Shavkat Butayev and Abbos Irisqulov (Butayev and Irisqulov, 2008), a collaborative result by the authors and experts at the Uzbek World Languages University and the Uzbek Academy of Sciences. The dictionary is one of the largest existing bilingual dictionaries available electronically and with one of the richest collection of entries.

The 2008 edition contains 40,000 lemmas in the English-Uzbek part, and about 30,000 is the Uzbek-English part. For each English word, it provides Uzbek senses in the following format.

Example 1 For the English word "sense", the dictionary stores the following information:

sense [sens] n **1)** *his, tuyg'u, sezgi;* **2)** *aql, fahm, idrok, zehn*,

where numbers represent each sense of "sense". ⊣

Remark 1 Each lemma's entry of (English-Uzbek part of) the dictionary contains the major parts of speech (PoSs) associated with the lemma. However, we shall see that our "connectivity restoration algorithm" (Section 6) uses only nouns, adjectives, verbs, and adverb, because it generates UZWORDNET from processing PWN and its semantic network. ⊣

5.2 Processing the dictionary

Now we provide some details on the preparatory tasks performed before running the main algorithm and presenting the human validators with resulting synsets. Here we focus on the first issue that we faced in processing the dictionary: the bad quality of the scan of the dictionary. It is worth mentioning that the electronic copy we used is an optical scan converted to text, which caused errors in parsing the dictionary for further use.

Example 2 Consider the entry in the dictionary:

abbey ['æbɪ] n **1)** *abbatlik...*

Automatic reading produced:

abbey ['reblj n **1)** *abbatlik...*

(closing bracket of the entry is misidentified as character "j"). ⊣

Character misinterpretations increased difficulty of applying parsing rules on the dictionary when converting it into more structured computable form for further use.

Specifically, to make the dictionary readable for the machine, individual pages were first enhanced visually and processed by a free OCR (Optical Character Recognition) service.[10] Successively, a series of complex regular expressions were written to parse individual translations from the dictionary and get rid of misinterpreted characters. Those were developed on the basis of observed erroneous patterns similar to the one described in Example 2.

5.2.1 Tabular format

The dictionary was converted into a convenient machine-readable form. In particular, we converted it into a table format where each row consisted of three columns: source lemma(s) (English); part of speech of source lemma(s); target lemma(s) (Uzbek translation by dictionary).

Example 3 The entry for *abbey* in the dictionary (source lemma) is converted into the following table format: <abbey; n; abbatlik, monastir>. ⊣

[10] Available at https://www.onlineocr.net.

Because of PWN'structure contains distinct database files for nouns, verbs, adverbs, and adjectives, the dictionary in tabular format was split into four separate files, one for each respective part of speech. The resulting four tabular dictionaries were sorted alphabetically by source lemma(s), in order to increase the speed of search for a particular lemma from PWN when it is used in the automatic construction of the wordnet.

6 Automatic Construction

The main procedure for building up UZWORDNET is an automatic translation—called "connectivity restoration algorithm" in reason of the most significant part of it (CRA; Algorithm 1)—of PWN (version 3.0) into Uzbek provided by the lexical resource (subsection 5.1) preprocessed into tabular format and files for each part of speech (subsection 5.2). The algorithm exploits the expansion method, as we accept the temporary assumption (see Future Work; Section 9) that the semantic structure of PWN is similar to the semantic structure of target language, the Northern Uzbek for us.

Algorithm 1: Connectivity Restoration.

Input : S, a data.pos file from Princeton WordNet (PWN, v3.0)
Input : D, English-Uzbek dictionary in tabular form for a specific PoS
Output: W, the UZWORDNET (UZW, v1.0)

```
1  W ← ∅
2  for each synset ∈ S do
3      for each lemma ∈ synset do
4          if lemma ∈ D then
5              W ← W ∪ translate(synset, D[lemma])
6  for each w_synset ∈ W do
7      if parent(w_synset) ∉ W then
8          s_synset ← parent(w_synset)
9          while s_synset ∉ W and s_synset ≠ top_level_synset(S) do
10             s_synset ← S[parent(s_synset)]
11             parent(w_synset) ← synset
12 return W
```

6.1 Connectivity Restoration Algorithm

UZWORDNET's development process is designed by the algorithm according to few related steps.

1. (lines 1-5): initial construction. The algorithm starts by initializing an empty set W for the resulting wordnet. English lemmas for each synset in S (file data.pos of PWN) are searched in D (dictionary in tabular format, cf. subsection 5.2.1). If a match is found, a new entry (Uzbek synset) in W is added. As the result, the algorithm produces the set W of synsets in Uzbek.

   However, not all synsets from PWN are translated into Uzbek. The reason is a lack of English entries in the lexical resource compared to the available lemmas from PWN. As a consequence, in W there may be disjoint synsets. This is the case of formation of *lexical gaps* for the target language, cf. (Giunchiglia et al., 2018; Giunchiglia et al., 2017), which means that the target language does not have, according to the lexical resource used, an equivalent synset.

2. (lines 6-12): connectivity restoration. For each synset from W (w_synset), the algorithm checks if the parent of w_synset exists in W. If it does not, then the algorithm extracts the parent of that synset from PWN (actually, from S in Algorithm 1) and checks if it exists in W. If not, it checks if the parent synset, say s, of that parent of w_synset exists in S. And so on until s is eventually found such that (a) s is translated into an Uzbek synset, say s', that is a (indirect) parent of w_synset, and (b) s' is in W. In the case that such a synset s satisfying conditions (a) and (b) above is not found, and the algorithm checked the synset from S, say s_r, that represents PWN's structural top level (root), then s_r becomes the parent of w_synset.

   As the result, all synsets in W are interconnected into the semantic hierarchy required.

Example 4 Consider Figure 5. Nodes denote synsets at a particular level in the structure; arrows denote the parental relationship between synsets.

The algorithm checks if an English synset, that is, a node from the structure of PWN, refers to a non-existent synset of Uzbek (target language) according to the lexical resource. In this case, we have a *lexical gap* for Uzbek language.

In the figure, S_D (synset S at level D in PWN) is referencing S_{C_2} (synset S at level C, child node 2 of parent node S_B in PWN). Assume that S_{C_2}

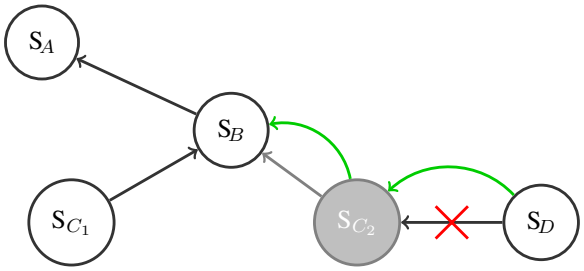

Figure 5: Visualization of the algorithm.

has no correspondent (equivalent) Uzbek synset, because the dictionary does not translate it. Then, S_{C_2} is not going to form a node of the emerging semantic network that eventually produced UZWORDNET. As a consequence, S_D references a synset (S_{C_2}) that is not represented in the resulting Uzbek wordnet. To avoid this problem, the algorithm searches out the parent node of S_{C_2} in the semantic structure of PWN and checks if it—node S_B in the figure—exists in the semantic structure that eventually builds the Uzbek wordnet.

For every synset s from PWN, the algorithm halts when run on s if either it finds a synset s' from PWN that is an *indirect* parent of s and s' has an equivalent Uzbek synset according to the dictionary, namely, s' exists in the Uzbek wordnet—like S_D and S_B, respectively, in Figure 5—or traversed the whole semantic network of PWN until it reached the synset at the root without finding a parental synset of s whose semantically equivalent Uzbek synset is provided by the dictionary—it would be the case of root node S_A that connects directly to S_D in Figure 5. ⊣

7 Expert Validation

Two native Uzbek speakers and one expert linguist—three co-authors of this paper—were asked to independently validate a sample of Uzbek synsets ("target lemmas") produced automatically.

For each part of speech (nouns, verbs, adverbs, and adjectives), an Excel file with 70 synsets ("lemmas") from PWN randomly selected was provided to the three "validators". For each file, the guidelines were the following:

1. Read each of the 70 English lemmas, its definition and example(s) if any.

2. For each lemma l, write 1 (meaning: "Yes, correct") in column "EVAL" you are provided for l, if you think that the target lemma, namely, Uzbek synset for l shares the same meaning of, or it is semantically equivalent to, l consistently with l's definition and, possibly, example(s). Write 0 ("No, wrong"), otherwise.

Some explanatory notes were also provided. In particular, this important one:

2.1 if English lemma l is translated into more than one word and *only some* of those words are the correct translation of l according to l's definition, *but some other words are not*, write 0 (i.e., "Translation incorrect").

8 Results

We run the CRA on PWN and the entries from the lexical resource as preprocessed (cf. subsections 5.2 and 5.2.1; also Input in Algorithm 1). The resulting semantic network, that is, UZWORDNET, contains 28140 synsets, 64389 sense and 20683 words, positioning UZWORDNET at the 18th place in the list of wordnets ranked by number of synsets, see Table 1 below (also cf. (Batsuren et al., 2019, Table 2).[11]

After the human evaluation over sample entries as described in the previous section, with a total number of instances processed be 17425 nouns, 5792 adjectives, 673 adverbs, and 4250 verbs, the estimated accuracy of the automatic translation by CRA resulted in 71.79% (Table 2).

8.1 Analysis

The estimated quality of UZWORDNET is rooted into a number of issues we encountered in processing the lexical resource. One important issue we mention here is strictly related to Uzbek rich semantics. Consider the following example.

Example 5 Suppose that our aim is to automatically extract from the dictionary the translation (synsets and senses, in particular) of the English word *body* stored in PWN and therein defined as follows: "The physical structure, including the bones, flesh, and organs, of a person or an animal".

Observe that, in the dictionary, *body*, as a noun, has the following Uzbek translations (senses):

1) *odam tanasi*; **2)** *so'zl. odam*; **3)** *murda*; **4)** *(nimaningdir) asosiy qismi*; **5)** *odamlar guruhi*.

Here is one example of sentence for each sense and its English translation (in parentheses):

[11] The list considers wordnets open source linked to PWN.

#	Language	Synsets	Senses	Words	Examples	Glosses	References
1	English	115424*	203145*	152059*	48459	109942	(Miller, 1995)
2	Finnish	107989	172755	115259	0	0	(Lindén et al., 2010)
3	Chinese	98324	123397	91898	17	541	(Wang and Bond, 2013)
4	Thailand	65664	83818	71760	0	0	(Thoongsup et al., 2009)
5	French	53588	90520	44485	0	0	(Sagot and Fišer, 2008)
6	Romanian	52716	80001	45656	0	0	(Tufiş et al., 2008)
7	Japanese	51366	151262	86574	28978	51363	(Bond et al., 2009)
8	Catalan	42256	66357	42444	2477	6576	(Gonzalez-Agirre et al., 2012)
9	Slovene	40233	67866	37522	0	0	(Fišer et al., 2012)
10	Portuguese	38609	60530	40619	0	0	(de Paiva et al., 2012)
11	Spanish	35232	53140	32129	651	17256	(Gonzalez-Agirre et al., 2012)
12	Polish	35083	87065	59882	0	0	(Piasecki et al., 2009)
13	Italian	33560	42381	29964	1934	2403	(Pianta et al., 2002)
14	Indonesian	31541	92390	24081	9	3380	(Noor et al., 2011)
15	Malay	31093	93293	23645	0	0	(Noor et al., 2011)
16	Basque	28848	48264	25676	0	0	(Pociello et al., 2011)
17	Dutch	28253	57706	40726	0	0	(Postma et al., 2016)
18	**Uzbek**	28140	64389	20683	0	0	this paper
19	Mongolian	23665	40944	26857	213	2976	(Batsuren et al., 2019)
20	Croatian	21302	45929	27161	0	0	(Oliver et al., 2016)

Table 1: Wordnets for number of synsets, cf. (Batsuren et al., 2019), *modified* (* our counting).

	Accuracy				
Validators	nouns	verbs	adverbs	adjectives	**Average**
MM	62.86 %	60.00 %	82.86 %	58.57 %	66.07%
NA*	78.57 %	71.43 %	84.29 %	72.86 %	*76.79 %*
UK	67.14 %	65.71 %	81.43 %	75.71 %	72.50 %
Average	69.52%	65.71 %	82.86%	69.05 %	**71.79 %**

Table 2: Human evaluation and accuracies (* expert linguist).

1) "*Faqatgina D va K vitaminlarini **odam tanasi** mustaqil ishlab chiqara oladi*". (The human body can only produce vitamins D and K.)

2) "***Odam** bu yerda yo'qolishi va hech qachon topilmasligi mumkin*". (A body could get lost out here and never be found.)

3) "***Murdalar** ertak aytmaydi*". (Dead men tell no tales.)

4) "*O'zbekistonda maoshlarning **asosiy qismi** oziq-ovqatga sarflanadi*". (In Uzbekistan a large part of salaries is spent on food.)

5) "*Bu **odamlar guruhi** o'zlarini xavf ostiga qo'yishmoqda*". (This group of people put themselves in danger.)

Further note that only the first translation, *odam tanasi*, matches PWN's definition of *body*.

However, our algorithm (CRA) extracts all five translations, even if we only need the senses of the source lemma that match the definition. ⊣

The example rises interesting questions about the semantic structure of UZWORDNET and polysemy. Although a deeper study into sense granularity in UZWORDNET and its effect on sense clustering is kept for future work, below we provide first answers and some further questions.[12]

8.2 Structure of UZWORDNET

Similarly to all word-nets created from PWN by expansion, nouns, verbs, adjectives and adverbs in UZWORDNET are grouped and classified into synonym sets (synsets), the major (lexical) relationship in the word-net. The semantic tree-like structure of synsets for nouns and verbs is based on the hypernym-hyponym relationship.

The structure for nouns, in particular, results the most representative among processed parts of speech, with 17425 nodes over the 28140 synsets total of the word-net produced. Its topologi-

[12]Thanks to the reviewer who asked some of the questions.

cal data, for instance the mean of distances of a node (synset) to the structure's root or, more precisely, the number of edges that connect consecutive nodes leading to the root—4.15, with standard deviation: 1.40—reveal a major downside of the structure, namely: its shallowness. In fact, UZWORDNET' structure contains many synsets with same sense high in the hyponym tree.

8.2.1 Polysemy

A main general issue in word-nets, which impacts on usability, is polysemy.

We quantified polysemy in the semantic structure of UZWORDNET for nouns (Figure 6) and verbs (Figure 7) by counting lemmas in synsets.

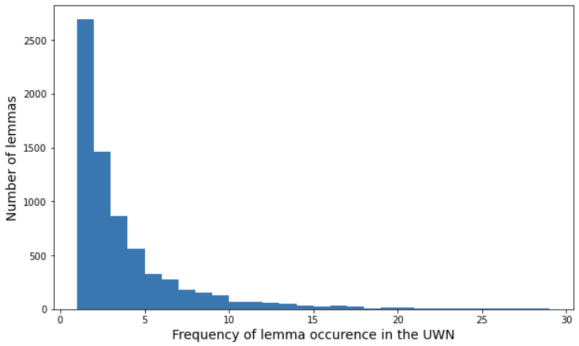

Figure 6: Degree of polysemy in nouns.

The mean of number of lemmas in synsets resulted in 2.05 (standard deviation: 3.56) for nouns and 2.99 (standard deviation: 4.78) for verbs.

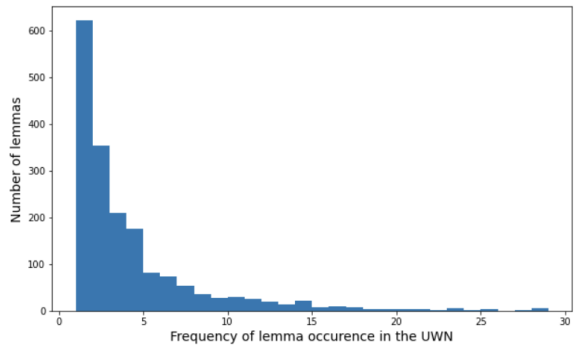

Figure 7: Degree of polysemy in verbs.

It turns out that polysemy is not present to a great degree on average in the structure. Moreover, most of the lemmas do not repeated more than mere several times in UZWORDNET.

A question is how polysemy and topological data we mentioned on average distance of nodes to UZWORDNET's root correlate. Precisely, the question here is: Where are synsets (for nouns, specifically) which contain more senses located on average within the semantic structure?

To assesses the degree of polysemy per level, namely, distance from root of the hypernym-hyponym tree, we run some scripts. The results are descriptive. Overall, the polysemy is higher closer to the root. An expected outcome, since senses closer to the root are more general, and therefore participating lemmas may express more concepts.

Remark 2 Another interesting question is: How much do sense granularity differ between the bilingual dictionary we used and UZWORDNET? Or, what senses for a given lemma in lexical resource we used are translated into senses for same lemma in UZWORDNET? An answer to the question, together with thoughtful analysis of sense granularity and sense coverage of UZWORDNET, would lead to interesting problems in the precision and recall of UZWORDNET and the issue of how to further improve it.[13] ⊣

Under a somewhat fortunate coincidence that most English lemmas in the dictionary we use do not have several senses in Uzbek, the issue of polysemy highlighted first by Example 5 and discussed further in this subsection could be solved by asking human experts to eliminate the translations automatically extracted that do not match the definition of the source lemma.

Although the extension of UZWORDNET by adding human expertise is out of the scope of this paper—and it is certainly part of future work, we like to foresee what would be the results.

8.3 Expert validation revised

We asked the validators to revise their validation over the identical set of sample files. The guidelines we gave to solve the task were identical to the previous, with the only exception over the explanatory note 2.1 (cf. Section 7).

The new explanatory note is:

2.1′ if English lemma l is translated into more than one word and *at least one* of those words is the correct translation of l according to l's definition, write 1 ("Translation correct").

The estimated accuracy of UZWORDNET after re-validation resulted in 75.98%. Table 3 reports in details the results of individual validations.

[13] For corpora to test our work on UZWORDNET upon coverage in words and senses, see for instance (Abdurakhmonova and Sobirov, 2019).

Validators	Accuracy (*revised*) nouns	verbs	adverbs	adjectives	Average
MM	74.29 %	67.14 %	87.14 %	77.14 %	76.43 %
NA*	75.71 %	71.43 %	84.29 %	67.14 %	*74.64 %*
UK	71.73 %	70.00 %	85.71 %	80.00 %	76.86 %
Average	73.91 %	69.52 %	85.71 %	74.76 %	**75.98 %**

Table 3: Human evaluation *revised* and accuracies (* expert linguist).

9 Conclusion and Future Work

In this paper, we have advanced and discussed the results on the initial development of UZWORDNET, a lexical-semantic database, or a "word-net", for the Northern Uzbek language compatible by expansion (extend/expansion method) to Princeton WordNet. UZWORDNET contains 28140 synsets, 64389 senses and 20683 words and is the output of an automatic process whose central procedure is an algorithm of connectivity run on Princeton WordNet' semantic network and an external lexical resource. Evaluation by three validators of UZWORDNET's accuracy in the translation, run over 280 sample entries, 70 for each PoS (nouns, verbs, adjectives, adverbs), resulted in an estimated accuracy of 71.79% minimum and 75.98% maximum according to the methodology of validation; 74.64% to 76.79% if considering only the evaluation by an expert linguist.

9.1 Future work

In the short term, we aim to make UZWORDNET available[14] among the Wordnets in the world, and to provide it open source under a license and format compatible with the Open Multilingual Wordnet (Bond and Paik, 2012; Bond and Foster, 2013) and other lexicgraphic data sets like Wikionary or other open source resources.[15] Moreover, to make UZWORDNET more accessible, we plan to build a simple SQL server and interface for using it. At the same time, we will refocus attention on our algorithms, improve the overall quality of automatic translation, and further investigate questions only addressed in this paper.

UZWORDNET has been developed by accepting the assumption that its semantic network is similar to the semantic structure of PWN. Obviously, it is *not* the case that Uzbek and English share exactly the same concepts, due to quite diverse underlying cultures of each language. Thus, we plan to keep the cultural diversity of Uzbek into more account. Before doing it, however, we plan to extend and improve UZWORDNET by expert human *translation* (for English lemmas not included in the lexical resource) or expert, selective *validation* (for English lemmas translated into more Uzbek synsets that need to be chosen according to definition; cf. Example 5), possibly using crowdsourcing (Ganbold et al., 2018; Fišer et al., 2014; Giunchiglia et al., 2015; Huertas-Migueláñez et al., 2018). We partially addressed to work to carry along this research direction and foresaw the results in subsections 8.1 and 8.3.

Successively, we aim to expand the core semantic structure of UZWORDNET to capture those features of the language that are typically Uzbek, that is, strictly and uniquely depending on Uzbek culture and not be available, as a consequence, in English-based PWN and other wordnets. In this way, both unicity and diversity of Uzbek language and, as a consequence, culture, will be modeled for the future use in IT applications. This extended version produced shall be *not* compatible to PWN (over concepts that are uniquely depending on Uzbek culture) and will be provided by working in partnership within the *DataScientia* initiative[16] using the Universal Knowledge Core (Giunchiglia et al., 2017; Giunchiglia et al., 2018), a multilingual, high quality, large scale, and diversity aware machine readable lexical resource.

Acknowledgments

The first author would like to thank Fausto Giunchiglia for proposing us to join *DataScientia*. Enver Menadjiev has supported our work in perspective to make UZWORDNET available online. Thanks to the anonymous reviewers, Alexandre Rademaker, Piek Vossen, Thierry Declerck, and all other participants to the conference for the interesting questions, comments and remarks.

[14] http://uzwordnet.ldkr.org/.

[15] About the format, we are evaluating to use XML or RDF formats, cf. https://globalwordnet.github.io/schemas/.

[16] http://datascientia.disi.unitn.it/.

References

[Abdurakhmonova and Khaydarov2019] Nilufar Abdurakhmonova and Muhammad Khaydarov. 2019. On the tasks of creating a Wordnet in the Uzbek language (uzbek). *O'zbekiston Xorijiy Tillar (Foreign Languages in Uzbekistan)*, 4:19–27. In Uzbek.

[Abdurakhmonova and Sobirov2019] Nilufar Abdurakhmonova and Abdulhay Sobirov. 2019. Korpus yordamida tezaurus yaratishning konseptual ahamiyati (Conceptual peculiarities on creation thesaurus by corpus). In *Proceedings of the International Conference on Translation, Information, Communication: Political and Social bridge*, pages 36–39. In Uzbek.

[Abdurakhmonova2020] Nilufar Abdurakhmonova. 2020. *Computational Linguistics*. Lambert Academic Publishing, Germany. In Uzbek.

[Abdurokhmonov and Darvishev2011] Sh. Abdurokhmonov and I. Darvishev. 2011. Workbook on the Uzbek Dialectology Course. `http://library.ziyonet.uz/ru/book/39716`, Namangan. In Uzbek.

[Batsuren et al.2019] Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, and Fausto Giunchiglia. 2019. Building the mongolian wordnet. In Christiane Fellbaum, Piek Vossen, Ewa Rudnicka, Marek Maziarz, and Maciej Piasecki, editors, *Proceedings of the Tenth Global WordNet Conference (GWC-2019)*, pages 238–244, Wrocaw, Poland. Oficyna Wydawnicza Politechniki Wrocawskiej.

[Bilgin et al.2004] Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. 2004. Building a Wordnet for Turkish. *Romanian Journal of Information Science and Technology*, 7(1-2):163–172.

[Bond and Foster2013] Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013, Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

[Bond and Paik2012] Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the Sixth Global WordNet Conference (GWC-2012)*, pages 64–71, Matsue, Japan. Global WordNet Association.

[Bond et al.2009] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 1–8.

[Boyd-Graber et al.2006] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In Petr Sojka, Key-Sun Choi, Christine Fellbaum, and Piek Vossen, editors, *Proceedings of the Third Global WordNet Conference (GWC-2006)*, pages 29–35, Brno, Czech Republic. Masaryk University.

[Butayev and Irisqulov2008] Shavkat Butayev and Abbos Irisqulov. 2008. *English-Uzbek, Uzbek-English Dictionary/Inglizcha-O'zbekcha, O'zbekcha-Inglizcha Lug'at*. O'zbekiston Respublikasi Fanlar Akad - Fan Nashriyoti, Tashkent, Uzbekistan.

[Çetinoğlu et al.2018] Özlem Çetinoğlu, Orhan Bilgin, and Kemal Oflazer. 2018. Turkish Wordnet. In Kemal Oflazer and Murat Saraçlar, editors, *Turkish Natural Language Processing. Theory and Applications of Natural Language Processing*, pages 317–336. Springer, Cham, Switzerland.

[de Paiva et al.2012] Valeria de Paiva, Rearden Commerce, and Alexandre Rademaker. 2012. Revisiting a brazilian wordnet. In *GWC 2012 6th International Global Wordnet Conference*, page 100.

[Eberhard et al.2020] David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. *Ethnologue: Languages of the World. Twenty-third edition*. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

[Ehsani et al.2018] Elin Ehsani, Ercan Solak, and Olcay Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):1–15, April.

[Ethnologue2020a] Ethnologue. 2020a. A Macrolanguage of Uzbekistan. `https://www.ethnologue.com/language/uzb`. Accessed: 2020-02-15.

[Ethnologue2020b] Ethnologue. 2020b. What are the Top 200 Most Spoken Languages? `https://www.ethnologue.com/guides/ethnologue200`. Accessed: 2020-02-15.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet—An Electronic Lexical Database*, Cambridge, MA. The MIT Press.

[Fišer et al.2012] Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. slownet 3.0: development, extension and cleaning. In *Proceedings of Sixth International Global Wordnet Conference (GWC 2012)*, pages 113–117.

[Fišer et al.2014] Darja Fišer, Aleš Tavčar, and Tomaž Erjavec. 2014. sloWCrowd: A crowdsourcing tool for lexicographic tasks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-14)*, pages 3471–3475, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Ganbold et al.2018] Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using crowd agreement for Wordnet localization. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-18)*, pages 474–478, Miyazaki, Japan. European Language Resources Association (ELRA).

[Giunchiglia et al.2015] Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláž ez, and Khuyagbaatar Batsuren. 2015. Crowdsourcing a large scale multilingual lexico-semantic resource. In Elizabeth Gerber and Panos Ipeirotis, editors, *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)*, Palo Alto, CA. AAAI Press.

[Giunchiglia et al.2017] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4009–4017, Menlo Park, CA. AAAI Press.

[Giunchiglia et al.2018] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed A. Freihat. 2018. One world - seven thousand languages. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CiCling-18)*, Hanoi, Vietnam.

[Gonzalez-Agirre et al.2012] Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, volume 2525, page 2529.

[Hirsch2005] Francine Hirsch. 2005. *Empire of nations: Ethnographic Knowledge & the Making of the Soviet Union*. Cornell University Press, Ithaca, N.Y.

[Huertas-Migueláž ez et al.2018] Mercedes Huertas-Migueláž ez, Natascia Leonardi, and Fausto Giunchiglia. 2018. Building a lexico-semantic resource collaboratively. In Jaka Čibej, Vojko Gorjanc, Iztok Kosem, and Simon Krek, editors, *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (ELX-18)*, pages 827–834, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

[Knight2016] Will Knight. 2016. AI's language problem. *MIT Technology Review*, (Sep/Oct).

[Lindén et al.2010] Krister Lindén, Lauri Carlson, et al. 2010. Finnwordnet-wordnet på finska via översättning. *LexicoNordica*.

[Matlatipov et al.2018] San'atbek Matlatipov, Mirsaid Aripov, and Nilufar Abdurakhmonova. 2018. Modeling WordNet type thesaurus for Uzbek language semantic dictionary. *International Journal of Systems Engineering*, 2(1):26–28.

[Miller et al.1990] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

[Miller1995] G. A. Miller. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Neale2018] Steven Neale. 2018. A survey on automatically-constructed wordnets and their evaluation: Lexical and word embedding-based approaches. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-18)*, pages 1705–1710, Miyazaki, Japan. European Language Resources Association (ELRA).

[Noor et al.2011] Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 255–264.

[Oliver et al.2016] Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2016. Automatic expansion of croatian wordnet. *Metodologija i primjena lingvističkih istraživanja*, page 171.

[Pianta et al.2002] Emanuele Pianta, Luids Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First Global WordNet Conference (GWC-2002)*, pages 293–302.

[Piasecki et al.2009] Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

[Pociello et al.2011] Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

[Postma et al.2016] Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the Eight Global Wordnet Conference, Bucharest, Romania*.

[Pulatov2011] Abdumajid Qayumovich Pulatov. 2011. *Computational Linguistics*. Akademnashr, Tashkent, Uzbekistan. In Uzbek.

[Rakhimov2011] Azamatjon Rakhimov. 2011. *The Foundations of Computational Linguistics*. Akademnashr, Tashkent, Uzbekistan. In Uzbek.

[Sagot and Fišer2008] Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In A. Oltramari, L. Prévot, C-R. Huang, P. Buitelaar, and P. Vossen, editors, *Proceedings of OntoLex (OntoLex-2008)*, pages 14–19.

[Steels et al.2002] L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. 2002. Crucial factors in the origins of word-meanings. In A. Wray, editor, *The Transition to Language*, pages 214–217. Oxford University Press, Oxford, UK.

[Steels1997] L. Steels. 1997. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

[Thoongsup et al.2009] Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 139–144.

[To'ychiboev and Khasanov2004] B. To'ychiboev and B. Khasanov. 2004. *Uzbek Dialectology*. Abdulla Qodiriy National Heritage.

[Tufiş et al.2008] Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian Wordnet: Current state, new applications and prospects. In Attila Tanács, Dora Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference (GWC-2008)*, pages 441–452, Szeged, Hungary. University of Szeged.

[Vossen1998] Piek Vossen. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer/Springer, Dordrecht, Netherlands.

[Wang and Bond2013] Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

[Wittgenstein1953] L. Wittgenstein. 1953. *Philosophical Investigations*. Blackwell, Oxford, UK.