Linking SIL Semantic Domains to Wordnet and Expanding the Abui Wordnet through Rapid Word Collection Methodology

Luis Morgado da Costa[1], **František Kratochvíl**[2], **George Saad**[2], **Benidiktus Delpada**[3], **Daniel Simon Lanma**[3], **Francis Bond**[2], **Natálie Wolfová**[2], and **A.L. Blake**[4]

[1]Vrije Universiteit Amsterdam, the Netherlands
[2]Palacký University Olomouc, Czech Republic
[3]Universitas Tribuana Kalabahi, Indonesia
[4]University of Hawai'i at Mānoa, USA

Abstract

In this paper we describe a new methodology to expand the Abui Wordnet through data collected using the Rapid Word Collection (RWC) method – based on SIL's Semantic Domains. Using a multilingual sense-intersection algorithm, we created a ranked list of concept suggestions for each domain, and then used the ranked list as a filter to link the Abui RWC data to wordnet. This used translations from both SIL's Semantic Domain's structure and example words, both available through SIL's Fieldworks software and the RWC project. We release both the new mapping of the SIL Semantic Domains to wordnet and an expansion of the Abui Wordnet.

1 Introduction

In this paper we describe the second phase of the Abui Wordnet construction which merges the data collected through the Rapid Word Collection method (RWC), as described in Section 2.2, into the Abui Wordnet v1.0 (see Section 1.1). The RWC method is built around the SIL Semantic Domains ontology, discussed in detail in Section 2.1. Much of the work discussed in this paper is related to the necessity of providing structure to data collected using common methods in Field Linguistics. The SIL Semantic Domains, and the Rapid Word Collection methodology in particular, support lexicographic work on endangered languages and significantly accelerate dictionary production. This paper looks into solving these issues by providing support to link unstructured types of data collected on the field to the Abui Wordnet.

1.1 Abui Wordnet

The Abui Wordnet was developed following the expansion approach (Kratochvíl and Morgado da Costa, 2022). Through a naive multilingual sense intersection algorithm, described in Section 3, we linked the data collected over the last two decades through the traditional descriptive workflow for which English, Indonesian, and Alor Malay glosses exist in the Abui dictionary (Kratochvíl and Delpada, 2014). The first version of the Abui Wordnet contained 1,475 synsets and 3,606 senses, and was entirely hand-checked by B. Delpada, who is a native speaker of Abui and one of the authors of this paper. This wordnet is released under the open CC-BY 4.0 license.[1]

2 Data Collection

In this section we provide an introduction to the structure and method for collecting our Abui data.

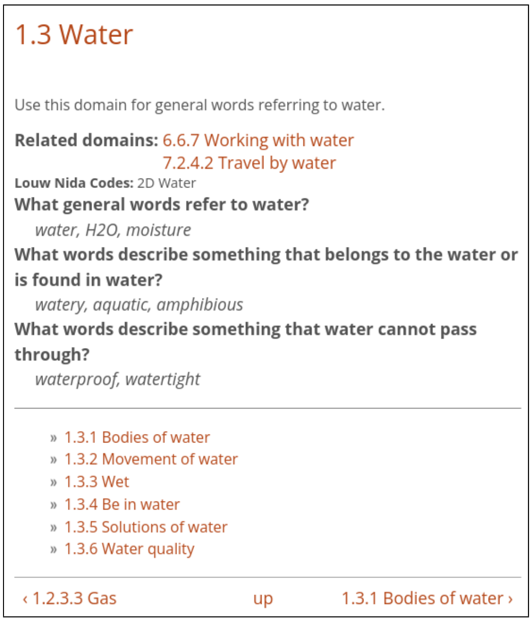
2.1 SIL Semantic Domains: Structure and Use

The SIL Semantic Domains[2] (SemDoms) is an ontology created by the Summer Institute of Linguistics linguist, Ronald Moe, to help investigate relationships among words. It builds on the long tradition of ontologies and thesauri developed in comparative linguistics and theology (see, e.g., Buck, 1949; Louw and Nida, 1992).

SemDoms are organized in an associative way, grouping words used to talk about a topic, regardless of their subtle differences. For example, as shown in Figure 1, the SemDom 1.3 Water is linked with two more SemDoms (6.6.7 Working with water and 7.2.4.2 Travel by water), which contain water-related action verbs. Each SemDom includes questions that elicit synonyms, such as *water*, H_2O, and *moisture*. The ontology also tracks associated properties such as *watery*, *aquatic*, or *amphibious*, and even loosely associated *waterproof* and *watertight*. Subdomains describe bodies of water, water movement, etc.

SemDoms facilitate dictionary building and have been incorporated and supported in various SIL

[1] https://github.com/fanacek/abuiwn
[2] http://semdom.org/



1.3 Water

Use this domain for general words referring to water.

Related domains: 6.6.7 Working with water
7.2.4.2 Travel by water

Louw Nida Codes: 2D Water

What general words refer to water?

water, H2O, moisture

What words describe something that belongs to the water or is found in water?

watery, aquatic, amphibious

What words describe something that water cannot pass through?

waterproof, watertight

- » 1.3.1 Bodies of water
- » 1.3.2 Movement of water
- » 1.3.3 Wet
- » 1.3.4 Be in water
- » 1.3.5 Solutions of water
- » 1.3.6 Water quality

‹ 1.2.3.3 Gas up 1.3.1 Bodies of water ›

Figure 1: SIL Semantic Domain for 1.3 Water

Languages	SemDom Titles	SemDom Words	Total
French	2,005	47,706	49,711
Spanish	2,056	45,801	47,857
English	2,013	41,494	43,507
Hindi*	2,202	34,544	36,746
Chinese	1,514	31,230	32,744
Portuguese	1,746	27,121	28,867
Indonesian	2,043	20,522	22,565
Nepalese*	2,061	17,770	19,831
Farsi	1,323	17,949	19,272
Urdu*	2,235	11,724	13,959
Bengali*	1,899	951	2,850
Russian*	2,673	3	2,676
Khmer*	2,120	0	2,120
Thai	1,555	1	1,556
Total	27,445	296,816	324,261

Table 1: SemDom data extracted from SIL FieldWorks, sorted by total number of data points per language; Data for Portuguese and Persian existed even though it was not properly advertised by SIL; Languages marked with * were missing from the OMW

software tools for language documentation, such as the SIL Toolbox, SIL Fieldworks (corpus, lexicon, parser), SIL Lexique Pro, and WeSay (dictionary).[3]

Multilingual versions of SIL Semantic Domains exist for 14 world languages,[4] including Chinese, French, Indonesian, Malay, Spanish, Swahili, and Urdu. However, not all translations are equally extensive, as shown in Table 1.

2.2 Rapid Word Collection Workshops

The Rapid Word Collection (RWC) method accelerates the lexicographic work by involving language communities, and has been used in over a hundred communities by untrained native speakers. It relies on a set of questions, derived from the SIL Semantic Domains, described above. The method exploits the brain's ability to rapidly recall words belonging to the same semantic domain. Speakers typically do not find this tiring and enjoy the process.

The questions are accompanied by answer sheets to record the semantic domain number, speaker details, and the vernacular word with their translations. Participants work in small groups or individually, according to their individual preference.

According to the RWC website,[5] two-week workshops consistently achieve 10,000 or more raw entries. This surpasses the 4,000 to 5,000 words collected over several years by a single language worker.

The RWC workflow yields a lexicon where most unique lexical entries have multiple senses, as is the case in dictionaries of resource-rich languages. The coverage is also not biased by a corpus, which is a big problem in the standard descriptive workflow. It is extremely difficult to reach the lexical breadth the RWC workshops can provide. Corpus-based methods are slow to elicit new words. One would need a corpus of upwards of one million words to collect a dictionary comparable to a two-week RWC workshop.

The RWC workshops demonstrate the wealth of lexical knowledge accumulated in minority languages and boost participants' confidence as well as language awareness. Realizing that some of the words may not be known by younger speakers, participants are challenged to assess the vitality of their language and their own commitment to promoting their language and culture. Finally, these workshops provide detailed information on community's orthographic preferences. A practical orthography may also be designed based on the RWC input.

2.3 RWC Workshops on Abui

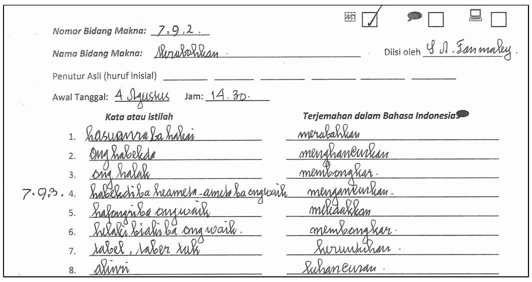
So far, we have held three RWC Workshops for Abui (in 2013, 2014, and 2016). In total they lasted 10 working days, with 25 people involved, on average, on any day. In total 67 Abui men and 21 Abui

[3] All available here: https://software.sil.org/

[4] https://rapidwords.net/resources (provides an incomplete list)

[5] https://rapidwords.net/

Figure 2: Abui participants of the Rapid Word Collection workshop, July 22-26, 2013



Nomor Bidang Makna: 7.9.2.

Nama Bidang Makna: Merubohkan. Diisi oleh S.A. Fanmaley.

Penutur Asli (huruf inisial) ______ ______ ______ ______ ______ ______

Awal Tanggal: 4 Agustus Jam: 14.30.

	Kata atau istilah	Terjemahan dalam Bahasa Indonesia
1.	hasuanra ba hakai	merobohkan
2.	ong habekda	menghancurkan
3.	ong halak	membongkar.
7.9.3. 4.	habekdi ba heamesa-amesa ba ong waik	mengancurkan.
5.	hafongri ba ong waik	meledakkan
6.	helaki kiali ba ong waik.	membongkar.
7.	tabel, taber tak	keruntuhan.
8.	alivri	kehancuran.

Figure 3: Rapid Word Collection worksheet example: domain 7.9.2 Tear down by S.A. Fanmaley

women participated, representing the Takalelang dialect and the adjacent areas, and contributing more than 17,000 raw entries. Figure 2 shows the Abui participants at work, writing or recording.

The participants recorded their answers on paper forms, indicating the Semantic Domain number, Abui words, and their Indonesian or Malay equivalents, as shown in Figure 3. Several Abui university students with adequate computer skills helped digitize the hand-written entries (including creating audio recordings and spreadsheets). This digitization work is still ongoing, with a small team working on the Indonesian and English translations, with about 12,300 words digitized to date.

3 Methodology

The work presented in this paper uses and extends the idea of Multilingual Sense Intersection (Bond et al., 2008; Bonansinga and Bond, 2016). The methodology is illustrated in Figure 4: it attempts to perform Word Sense Disambiguation (WSD) — i.e., to determine the most likely sense of a word with reference, e.g., to a wordnet hierarchy — by restricting the available semantic space through the intersection of semantic spaces of aligned translations of that same word. This method has been used to create new wordnets, such as the Coptic Wordnet (Slaughter et al., 2019) and, most recently, also to kick-start the development of the Abui Wordnet from field data (Kratochvíl and Morgado da Costa, 2022).

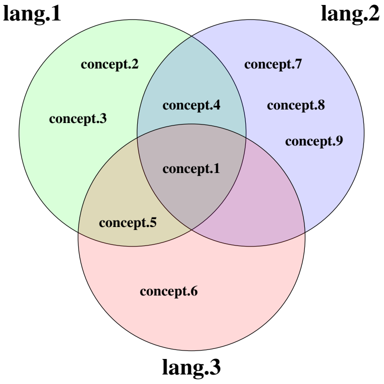


Figure 4: Sense Intersection visualization: each colored circle represent a different language (lang.1-3); concepts (concept.1-9) represent the ambiguity of a single lemma within that language; The higher the number of languages, the smaller the intersected space – yielding fewer and fewer sense candidates;

In this work, we employ this same concept in two ways: i) we use Multilingual Sense Intersection to perform WSD to map the SIL SemDoms data to the Open Multilingual Wordnet (OMW, Bond and Foster, 2013); ii) we use the results of the previous step as a pivot to map Abui data collected through RWC Workshops to the Abui Wordnet.

3.1 Linking SIL Semantic Domains to OMW

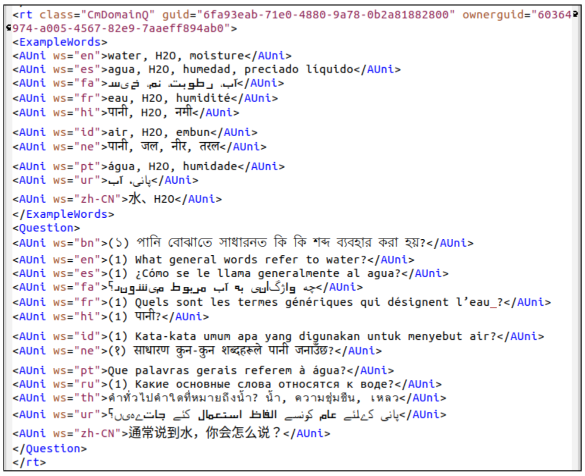
The idea of linking SIL SemDoms to wordnet was first proposed by Rosman et al. (2014).

As discussed in Section 2.1, SemDoms are mostly used for language documentation. To this end, there has been a considerable community effort to translate this resource. Translations of this resource are most commonly released as localization packages[6] for SIL FieldWorks[7] – an open-source project designed to help collect and publish dictionary data, including support dictionary development through SemDoms. It also supports interlinearization of texts and morphological analysis.

Our primary goal to link SemDoms to OMW was to be able to pivot this information to improve our ability to better link the Abui data collected using the RWC Workshop method, described in Section 2.3. To achieve this, we wanted to link not only the SemDom titles (as referred within FieldWords)

[6] https://software.sil.org/fieldworks/download/localizations/

[7] https://github.com/sillsdev/FieldWorks



```
<rt class="CmDomainQ" guid="6fa93eab-71e0-4880-9a78-0b2a81882800" ownerguid="60364
974-a005-4567-82e9-7aaeff894ab0">
<ExampleWords>
<AUni ws="en">water, H2O, moisture</AUni>
<AUni ws="es">agua, H2O, humedad, preciado líquido</AUni>
<AUni ws="fa">آب، رطوبت، نم، خیس</AUni>
<AUni ws="fr">eau, H2O, humidité</AUni>
<AUni ws="hi">पानी, H2O, नमी</AUni>
<AUni ws="id">air, H2O, embun</AUni>
<AUni ws="ne">पानी, जल, नीर, तरल</AUni>
<AUni ws="pt">água, H2O, humidade</AUni>
<AUni ws="ur">پانی، آب</AUni>
<AUni ws="zh-CN">水、H2O</AUni>
</ExampleWords>
<Question>
<AUni ws="bn">(১) পানি বোঝাতে সাধারনত কি কি শব্দ ব্যবহার করা হয়?</AUni>
<AUni ws="en">(1) What general words refer to water?</AUni>
<AUni ws="es">(1) ¿Cómo se le llama generalmente al agua?</AUni>
<AUni ws="fa">چه واژگانی به آب مربوط می‌شوند؟</AUni>
<AUni ws="fr">(1) Quels sont les termes génériques qui désignent l’eau_?</AUni>
<AUni ws="hi">(1) पानी?</AUni>
<AUni ws="id">(1) Kata-kata umum apa yang digunakan untuk menyebut air?</AUni>
<AUni ws="ne">(१) साधारण कुन-कुन शब्दहरूले पानी जनाउँछ?</AUni>
<AUni ws="pt">Que palavras gerais referem à água?</AUni>
<AUni ws="ru">(1) Какие основные слова относятся к воде?</AUni>
<AUni ws="th">คำทั่วไปคำใดที่หมายถึงน้ำ? น้ำ, ความชุ่มชื้น, เหลว</AUni>
<AUni ws="ur">پانی کےلئے عام کونسے الفاظ استعمال کئے جاتےہیں؟</AUni>
<AUni ws="zh-CN">通常说到水，你会怎么说？</AUni>
</Question>
</rt>
```

Figure 5: XML data extracted from SIL Fieldworks for the first question of SemDom 1.3 Water

but also – and most importantly – we wanted to link the answers to the questions within each SemDom (see Figure 1), as this is the primary data collected during the RWC workshops. These words are referred to as example words within FieldWorks, so this is the nomenclature we will use.

This is one of the main differences between our work and Rosman et al. (2014). In their work, only SemDom titles are linked to Wordnet. For our goals, this would not be enough. As noted by Rosman et al. (2014), even relations between domains and their subdomains are not typed in the same way as you would find in wordnets – making reference to the so-called ‘Tennis Problem’, which describes the fact that wordnets do not link clearly related words such as *tennis*, *racket*, *ball*, and *net*. This problem would most certainly be exacerbated when considering the relation between SemDom titles and example words – which are the basis of the RWC method.

The second main difference with the previous work mapping SemDoms was the number of languages used to attempt the mapping. While the previous work only had access to English and Indonesian data at the time of publication, we had access to a much larger collection of languages.

We created a new project on FieldWorks and imported all languages known to contain translations for SemDoms (including only partial translations). This generated an XML file containing parallel data in all available languages. This data is split into data concerning the SemDom titles, and data concerning questions and answers within a SemDom. Figure 5 shows an example of how the data is organized for the first question of the SemDom 1.3 Water – see also Figure 1, for reference.

The results of this data extraction were summarized in Table 1. In total, we extracted over 324,000 expressions (including words and multiword-expressions), about 8.5% of which were related to SemDom titles, and the remainder to SemDom example words. We were able to extract data for 14 languages. French was the language with most words, followed by Spanish, English, Hindi and Chinese. Some languages only had a partial translation for the full SemDom hierarchy. The reason why some languages seem to have more titles than there are SemDoms is due to the fact that some semantic domains actually include a list of words in their title (e.g. SemDom 1.5.4 Moss, fungus, algae). Both titles and example words were split on commas (which were different Unicode characters for different languages).

This data was then mapped to wordnet using the data from the OMW (Bond and Foster, 2013).

3.2 Expanding the OMW

The Open Multilingual Wordnet (OMW 1.0: Bond and Foster, 2013) links dozens of open wordnets projects in a massively multilingual database, using the Princeton WordNet (Fellbaum, 1998) as the pivot structure.

Fortunately, in addition to the English Princeton WordNet, the OMW already included wordnet projects for many of the necessary languages, including: WOLF (Sagot and Fišer, 2008), for French; the Multilingual Central Repository (Gonzalez-Agirre et al., 2012), for Spanish; the Chinese Wordnet (Huang et al., 2010) and the Chinese Open Wordnet (Wang and Bond, 2013), for Mandarin Chinese; the OpenWordnet-PT (de Paiva and Rademaker, 2012), for Portuguese; the Wordnet Bahasa (Mohamed Noor et al., 2011), for Indonesian and Malay; the Persian Wordnet (Montazery and Faili, 2010), for Farsi; and the Thai Wordnet (Thoongsup et al., 2009).

The OMW was missing data for Hindi, Nepalese, Urdu, Bengali, Russian and Khmer. For Khmer, even through there are reports detailing the construction of the Khmer WordNet (Phon and Pluempitiwiriyawej, 2020), we were not able to find or access the data. Since there were no SemDom example words for this language, we decided it was not worth pursuing it further.

For the remaining languages, there were actually wordnets being actively maintained, but they were not part of the OMW due to their restrictive licensing constraints (i.e. NonCommercial). Since this did not impede our work, we added the missing languages to our own local copy of the OMW.

Data for Hindi, Nepalese, Urdu and Bengali was provided by the IndoWordnet (Bhattacharyya, 2010), and its IndoWordnet-English Wordnet Mapping (Kanojia et al., 2018). Data for Russian was provided by the Russian Wordnet (Loukachevitch et al., 2016), which also includes a mapping to the PWN (Loukachevitch and Gerasimova, 2019).

It is also important to note that, for the work presented here, we used an extended version of the OMW which includes additions to the PWN's hierarchy through the annotation of the NTU-MC sense-tagged corpus (Tan and Bond, 2014; Bond et al., 2013; Wang and Bond, 2014; Bond et al., 2021), as well as other extensions including pronouns (Seah and Bond, 2014) and exclamatives (Morgado da Costa and Bond, 2016). As a result, our released data contains some offsets that do not directly map to the PWN.

With all the data in a single repository, the expanded OMW was used to map the SIL SemDom data using the method described above – multilingual sense intersection. The results of this experiment are discussed in Section 4.1.

3.3 Linking RWC data to the Abui Wordnet

After creating the mapping between the SemDoms and the OMW, we used it to help us link the Abui RWC data to the Abui Wordnet.

Even though previous work on the Abui Wordnet showed promising results using word sense intersection to find candidate senses, this method presupposed the data was provided (at least) in three languages. The problem with the Rapid Word Collection data was that we had only a limited number of translations for each Abui word.

For the 12,331 Abui words digitized to date, 12,324 words were translated to Indonesian, 9,078 words were translated to Alor Malay, and only 5,846 were translated to English. About 11,000 words were used for the linking described in Section 4.2, the additional 1,300 words were digitized since. However, each Abui word was also linked to the SemDom identifier that prompted the native speaker to provide that word. SemDoms were used at the level of identifier (i.e., they were not linked to a specific question within that identifier).

What we wanted to verify was if, after properly linking the SemDoms to the OMW, we could use this mapping to further filter the data provided by RWC. To do this, first, we performed multilingual sense intersection using only the data provided through the RWC method, as described in Section 4.1. We then used the data provided by our SemDom mapping (with different kinds of confidence level), to check if intersecting these two mappings could be used to reliably increase the quality of new senses suggested for the Abui Wordnet. The results for these experiments are detailed in Section 4.2.

4 Results and Evaluation

In this section we discuss three different things: i) the results of mapping the SemDoms to the OMW using the data extracted from SIL FieldWorks; ii) the results of producing sense candidates for the Abui Wordnet through multilingual sense intersection using the RWC method with and without using i) as as filtering step; and iii) the results of hand-checking sense candidates produced in ii) by a group of linguists and native speakers.

4.1 Mapping SIL Semantic Domains to OMW

Using the data presented in Table 1, we extracted data for all 1,792 different SemDom identifiers. Using the method briefly described in Section 3, we performed multilingual sense intersection for each level of the SemDom hierarchy. However, we split this intersection into two parts: i) using data pertaining only to SemDom titles; and ii) using data pertaining only to SemDom example words (i.e., answers to the questions in that SemDom).

The reason to separate these two sets of data is quite intuitive. For i), we are trying to link the actual SemDom to the OMW. While this could well be a many-to-one mapping (i.e., many wordnet senses mapped to a single SemDom), there is a finite/correct set of links that should be made between these two resources. For ii), however, this is not true. The large majority of SemDom questions are open ended (e.g., 'What utensils are used to cut food?', from SemDom 5.2.1.3 Cooking utensil). The work of translating the SemDom is not strictly to translate example words that have been included in previous languages, and people are welcome to include more/different examples. We have noticed, for example, that both French and Spanish go well beyond the list of words provided for English (the original language).

Intersected Languages	SemDom Titles	SemDom Words
1 lang	29,986	293,821
2 langs	6,233	58,320
3 langs	2,524	23,074
4 langs	1,355	10,782
5 langs	804	5,595
6 langs	466	2,403
7 langs	267	317
8 langs	108	-
9 langs	8	-
Total	41,751	394,312
>3 langs	5,532	42,171

Table 2: Number of candidate concepts for the mapping SIL SemDoms to OMW, organized by number languages suggesting each candidate

The results for the intersection experiments are summarized in Table 2. We provide the number of candidate concepts, sorted by number of languages intersected. Using any number of intersected languages, we collected about 41,700 candidates from SemDom titles, and about 394,000 candidates for SemDom example words. Some of these candidate concepts were suggested by as many as nine languages, although the large majority was suggested by either one or two languages.

However, we know from previous work that quality really spikes at a minimum of three intersected languages. Slaughter et al. (2019) reported that senses triangulated by three or more languages were shown to be correct as high as 98% of the time. Similarly, Kratochvíl and Morgado da Costa (2022) reported 99% accuracy for senses suggested by intersecting three languages, when building the Abui Wordnet. For this reason, we pruned the results of our mapping to only those provided by the intersection of three or more languages.

Our pruned results yielded over 5,500 OMW concepts linked to SemDom titles, and over 42,000 concepts linked to SemDom example words. These numbers are distributed over 1,173 SemDom titles with at least one link to OMW, and over 1,671 SemDom identifiers with at least one example word linked to OMW. We did not expect to provide mappings to all 1,792 SemDom identifiers. This is because because many SemDom titles and example words are, in fact, phrases and not words (e.g. SemDom 2.5.6 Symptom of disease, or 5.8 Manage a house). The fact that some SemDom idenfitiers did not link to OMW is a good sign of quality.

Intersected Langs.	Candidate Senses
1 lang	75,188
2 langs	5,065
3 langs	1
Total	80,254

Table 3: Number of sense candidates generated by the data collected using the RWC method

	SemDom 3 langs	SemDom 4-5 langs	SemDom >5 langs	Total
RWC 1 lang	4,821	4,146	1,048	10,015
RWC 2 langs	282	333	150	765
Total	5,103	4,479	1,198	10,780

Table 4: Number of sense candidates generated by the data collected using the RWC method after applying the filtering step of belonging to the SemDom mappings

4.2 Linking RWC data to the Abui Wordnet

In order to link the Abui data gathered from the RWC method, we started by performing sense intersection on the existing data. The results of this intersection is shown in Table 3.

As mentioned in Section 3.3, this data comprised about 11,000 Abui words, almost fully translated into Indonesian, but only partially translated into Malay and English. This resulted in a very limited ability to generate high levels of intersection. As shown in Table 3, only a single word was intersected by three languages.[8]

We knew from previous work that two-way intersection yields an accuracy of about 50%. While arguably useful, this score was lower than what we wanted to work with. The way we proposed to raise the confidence score was to use the mapping between SemDom titles and example words to OMW as a filter for the data presented in Table 3.

Since every Abui word collected through the RWC method was linked to a SemDom identifier, we were able to exclude senses that had not been predicted as likely members of that SemDom identifier using the mappings we created. We used the mappings for both the SemDom titles and the example words. This greatly reduced the number of candidate senses. A summary of the results after this filtering step can be see in Table 4.

As mentioned previously, the final mappings between SemDom titles and example words contained only concepts suggested by the intersection of three or more languages. In total, after the filtering step, 10,780 candidate senses remained. However, Ta-

[8] This word was the word for 'yes'.

ble 4 shows a more in-depth distribution of the data. In total the data was distributed into six groups divided into two axes: i) whether the sense candidate was suggested by one or by two languages during the intersection of the RWC data; and ii) whether the SemDom mappings had been suggested by the intersection of three languages, either four or five languages, or by more than five languages.

In general we assumed that the higher the intersection level of both axes, the higher the quality of the suggested senses. While this was not strictly true, hand-checking part of this data confirmed that our method was quite promising.

4.3 Hand-Checked Evaluation

Following the discussion for Table 3, above, we decided to hand-check a portion of each of the six classes of candidate senses, as listed in Table 4. The checking was performed by B. Delpada and D. Lanma (Abui native speakers and linguists) and F. Kratochvíl and G. Saad (linguists working on Abui). We believe that the use of this evaluation is two-fold: i) it directly evaluates the quality of candidate senses for the Abui Wordnet; and ii) it indirectly evaluates the quality of the SemDom mappings, because all candidate senses were filtered by this mapping.

We decided to hand-check 250 candidate senses from each of the six groups discussed above.[9] The results of this evaluation are provided in Table 5.

As is shown, all six groups show a fairly high accuracy of between 87.6% and 99.6%. We had assumed that the higher the intersection level of both axes, the higher the quality of the suggested senses. However, even though the data doesn't fully confirm our assumption, we believe we know why this happened. It has to do, in great part, with the quality of the Wordnet Bahasa – which contains data for both Indonesian and Malay (developed in parallel), two of the three languages contained in the Abui RWC data.

The sense candidates generated for RWC data intersection by two languages was quite limited. And it so happened that among the candidate senses for the groups with lowest accuracy were Abui words translated with words that contained a lot of incorrect data in these two wordnets. Since these two languages are very closely related, and the Wordnet Bahasa used the same methods to develop both languages, some of these errors have a bigger impact than they should.

One simple example to illustrate this problem is the Abui word ‘bilengra’, which has been glossed with the word ‘melukis’ for Indonesian and ‘draw’ for English. The problem that follows is that the lemma ‘melukis’ has 26 senses in the Wordnet Bahasa (Indonesian). Many of these senses are, in fact, incorrect. KBBI defined ‘melukis’ as a verb with the gloss ‘make drawings using pencils, pens, brushes, and so on, whether with color or not’.[10] However, in the Wordnet Bahasa, this lemma includes senses glossed as ‘bring, take, or pull out of a container or from under a cover’ (01995211-v), ‘suck in or take (air)’ (01199009-v), ‘cause to move in a certain direction by exerting a force upon, either physically or in an abstract sense’ (02103162-v) or ‘take liquid out of a container or well’ (01854132-v). It is not surprising that the PWN (correctly) adds the lemma ‘draw’ to all these concepts, hinting at why the Wordnet Bahasa may have included these incorrect senses, and showing the limitations of automatically built wordnets without incorporating a strong review cycle.

Despite some of these limitations, we are satisfied with the results we have achieved. The methodology we developed is robust enough to deal even with somewhat noisy data.

For the future, however, it is important to note that both Indonesian and Malay are essential languages in the production of language resources for Abui, since these are a few of the only other languages speakers of Abui can speak fluently. As such, working towards the improvement and maintenance of the Wordnet Bahasa is well in the interest of the Abui Wordnet and other minority languages of Indonesia.

5 Release Notes

This paper releases two new sets of data: i) the mapping of SIL Semantic Domains to OMW (through PWN 3.0 offsets); and ii) a new extension to the Abui Wordnet.

The mapping of SIL Semantic Domains to OMW will be shared under a Creative Commons Attribution-ShareAlike 4.0 license (following the original license for this resource). In the future, we will attempt to liaise with SIL and open the license further. This data will be released as two TSV files, one for the SemDom identifier titles, and another

[9] Except where mentioned in the table.

[10] Free translation from https://kbbi.kemdikbud.go.id/entri/melukis

	SemDom 3 langs	SemDom 4-5 langs	SemDom >5 langs
RWC 1 lang	0.956♣	0.952	0.996
RWC 2 langs	0.876*	0.932	0.913*

Table 5: Shows the accuracy of the matches, based on a sampled section of the data comprising 250 senses per condition (except cells marked with * for which all suggested senses were checked, see Table 4); After this initial evaluation (of 250 candidate senses per condition), and before the camera-ready version of this paper was submitted, all 4,821 members of the class marked with ♣ were hand-checked, yielding an updated accuracy score of 0.964 (i.e., higher than initially predicted)

file for example words within each identifier. The files contain the following information: SemDom identifier, suggested PWN 3.0 offset, number of languages intersected for this suggestion, and list of language names. This new dataset will be made available on GitHub.[11]

The second set of data concerns new sense candidates for the Abui Wordnet. Interestingly enough, of the 10,780 newly generated candidate senses, only 248 already existed in the Abui Wordnet. All data that has been hand-checked will be included in future version of the Abui Wordnet. The remainder of the data will also be released as separate files and incorporated into the Wordnet after it has been hand-checked (see Section 6). Both sets of data will be released in the existing Abui Wordnet GitHub repository,[12] and released under this wordnet's license – Creative Commons Attribution 4.0 International License.[13]

6 Future Work

This paper presents one of many steps towards the improvement of the Abui Wordnet and the wordnet infrastructure in general.

A natural next step is to finish hand-checking the list of candidate senses generated in this paper. Our hand-checking evaluation has checked 1,432 out of the existing 10,780 generated senses – leaving around 9,300 candidate senses that need to be checked. We hope to be able to do this with the help of the Abui community in the very near future.

Another natural step is to find ways to work with SIL directly and to produce a hand-checked mapping of SemDoms (titles and example words) to OMW. Once this is done, the multilingual nature of OMW could be used to produce official language translations for all available languages in the OMW – centralizing and accelerating the work that is now performed by individual groups of translators, for each language. If properly linked to the OMW, the PWN's semantic hierarchy could even be used to slightly expand the SemDoms by adding new example words to certain semantic classes that are well encoded in PWN's semantic hierarchy (e.g. animals, trees, professions, etc.).

An interesting idea we would like to pursue further is to push the sense-intersection one step forward and start investigating which languages yield best results when intersected. While the underlying idea that the more languages the better the candidate senses produced will undoubtedly hold truth, the quality of candidate senses produced by a 2-way or 3-way intersection may depend highly on which languages are involved. Languages that are closely related, such as Spanish and Portuguese, will arguably share more non-literal meaning extensions than other pairs of less related languages such as Spanish and Chinese. We believe that exploring our intersection methodology using languages from different families or languages that do not share a lot of their cultural background could be a great start for this future research direction.

Finally, we would like to exploit the mappings we provide for SemDom titles and example words to enrich the semantic hierarchy of wordnet projects. We believe that the association-based methodology inherent to SemDoms (and successfully exploited by the RWC method) is directly related to Common Sense Reasoning. Currently, the wordnet hierarchy is known to be both too fine-grained (Hayashi, 2022) and also lacking sufficient semantic relations (Di Caro and Boella, 2016) for tasks involving Common Sense reasoning. We believe our work mapping SemDoms to the OMW could be a good start for a project looking into these two issues.

7 Conclusion

In this paper we have used the idea of multilingual sense intersection for two ends: i) to create a new language resource – a mapping of SIL Semantic Domains to the structure of the Open Multilingual Wordnet; and ii) to use this new semantic resource

[11] https://github.com/lmorgadodacosta/sil-semantic-domains-wordnet-mapping

[12] https://github.com/fanacek/abuiwn

[13] https://creativecommons.org/licenses/by/4.0/

as a filter to expand the Abui Wordnet with data collected using the Rapid Word Collection method (which relies on the SIL Semantic Domains).

We have yielded very positive results for both goals. We have linked more than 47,500 OMW concepts to the SIL Semantic Domains (with a high confidence score), and we have generated more than 10,500 new sense candidates for the Abui Wordnet. Human evaluation has offered a confidence score for these sense candidates between 87.6% and 99.6%.

We hope our work inspires other linguists with data linked to SIL Semantic Domains to follow in our footsteps and to link their data to structures such as the OMW. We hope that lexicographic work on low-resource languages may benefit from both the OMW structure and the SIL experience in rapid lexicographic work involving language communities.

Acknowledgments

The authors acknowledge the generous support of the Czech Science Foundation grant 20-18407S Verb Class Analysis Accelerator for Low- Resource Languages - RoboCorp (PI Kratochvíl) and the EU's Horizon 2020 Marie Skłodowska-Curie grant H2020-MSCA-IF-2020 CHILL – No.101028782 (PI Morgado da Costa).

The authors contributed to the paper in the following way: i) Kratochvíl, Bond, Delpada, Lanma, and Saad conducted the Rapid Word Collection workshops and digitised the data, provided translations into Malay, Indonesian and English (Wolfová contributed many English translations); ii) Blake contributed an extensive ethnobotanical dataset to the database; iii) Morgado da Costa was responsible for extracting the data from SIL FieldWorks, implementing the sense intersection algorithm that produced the SIL mappings to OMW and a new list of sense candidates for the Abui Wordnet and, together with Kratochvíl, was the lead writer of this manuscript.

References

Pushpak Bhattacharyya. 2010. Indowordnet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proc. of the 8th Global WordNet Conference*, pages 44–49.

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

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Francis Bond, Andrew Kirkrose Devadason, Rui Lin Melissa Teo, and Luis Morgado Da Costa. 2021. Teaching through tagging — interactive lexical semantics. In *Proceedings of the 11th Global WordNet Conference (GWC 2021)*, Pretoria, South Africa. Global Wordnet Association.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 149–158.

Carl Darling Buck. 1949. *A dictionary of selected synonyms in the principal Indo-European languages : a contribution to the history of ideas*. Chicago University Press, Chicago.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Luigi Di Caro and Guido Boella. 2016. Automatic enrichment of wordnet with common-sense knowledge. In *10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 819–822. European Language Resources Association (ELRA).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Yoshihiko Hayashi. 2022. Towards the detection of a semantic gap in the chain of commonsense knowledge triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3984–3993.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).

Diptesh Kanojia, Kevin Patel, and Pushpak Bhattacharyya. 2018. Indian Language Wordnets and their Linkages with Princeton WordNet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

František Kratochvíl and Benidiktus Delpada. 2014. Abui-English-Indonesian Dictionary. 2nd. edition.

František Kratochvíl and Luís Morgado da Costa. 2022. Abui Wordnet: Using a toolbox dictionary to develop a wordnet for a low-resource language. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 54–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Natalia Loukachevitch and Anastasia Gerasimova. 2019. Linking Russian Wordnet RuWordNet to WordNet. In *Proceedings of the 10th Global Wordnet Conference*, pages 64–71, Wroclaw, Poland. Global Wordnet Association.

Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova, Vladimir V Ivanov, and Boris V Dobrov. 2016. Creating russian wordnet by conversion. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue*, pages 405–415.

Johannes P Louw and Eugene Albert Nida. 1992. *Lexical Semantics of the Greek New Testament: A Supplement to the Greek-English Lexicon of the New Testament Based on Semantic Domains*, volume Resources for Biblical Study of *Society of Biblical Literature*. Scholars Press, Atlanta.

Nurril Hirfana 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 (PACLIC 25)*, pages 258–267, Singapore.

Mortaza Montazery and Heshaam Faili. 2010. Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, pages 846–850.

Luís Morgado da Costa and Francis Bond. 2016. Wow! What a useful extension! Introducing non-referential concepts to WordNet. In *Proceedings of the 10th edition of the International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4323–4328, Portorož, Slovenia.

Udorm Phon and Charnyote Pluempitiwiriyawej. 2020. Khmer wordnet construction. In *2020 - 5th International Conference on Information Technology (InCIT)*, pages 122–127.

Muhammad Zulhelmy bin Mohd Rosman, František Kratochvíl, and Francis Bond. 2014. Bringing together over- and under- represented languages: Linking WordNet to the SIL Semantic Domains. In *Proceedings of the Seventh Global Wordnet Conference*, pages 40–48, Tartu, Estonia. University of Tartu Press.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 82–88.

Laura Slaughter, Luis Morgado Da Costa, So Miyagawa, Marco Büchler, Amir Zeldes, Hugo Lundhaug, and Heike Behlmer. 2019. The Making of Coptic Wordnet. In *Proceedings of the 10th Global WordNet Conference (GWC 2019)*, Wroclaw, Poland.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89.

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), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),*, Suntec, Singapore.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Sixth International Joint Conference on Natural Language Processing*, pages 10–18.

Shan Wang and Francis Bond. 2014. Building the sense-tagged multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).