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

Luis Morgado da Costa, František Kratochvíl, George Saad, Benidiktus Delpada, Daniel Simon Lanma, Francis Bond, Natálie Wolfová, and A.L. Blake

VU VRIJE UNIVERSITEIT AMSTERDAM

Palacký University Olomouc

UNIVERSITAS TRIBUANA KALABAHI

UNIVERSITY of HAWAI'I MĀNOA

European Commission Marie Skłodowska-Curie Actions

Goals

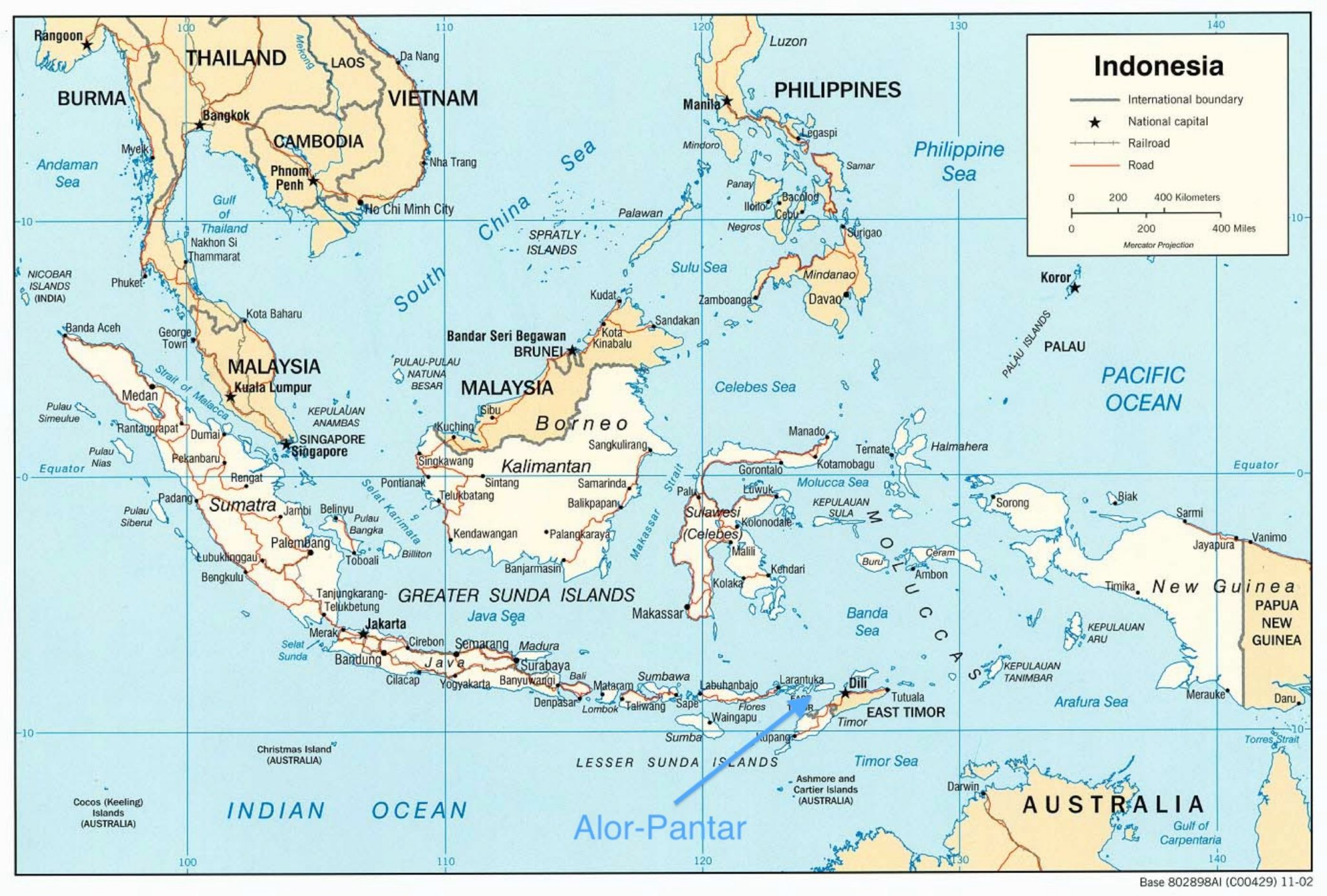
This is one step towards the larger goal of **inviting Field Linguists to consider wordnets as an interesting framework** to organize field data – respecting/celebrating linguistic diversity and helping low-resourced languages.

In this paper we describe **a new methodology to expand the Abui Wordnet** through data collected using the **Rapid Word Collection (RWC) method**.

In the process we ended up linking the **SIL's Semantic Domains to OMW** – using a naive **Multilingual Sense Intersection** algorithm.

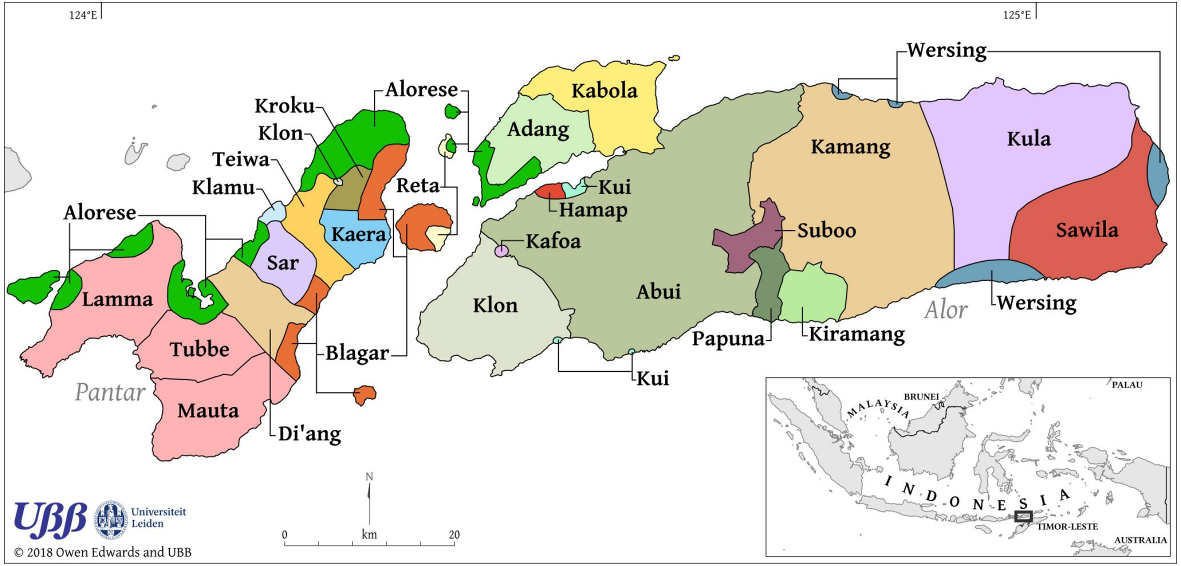
We release both the **new mapping of the SIL Semantic Domains** to wordnet and an **expansion of the Abui Wordnet**.

Abui (Alor Island, Eastern Indonesia)

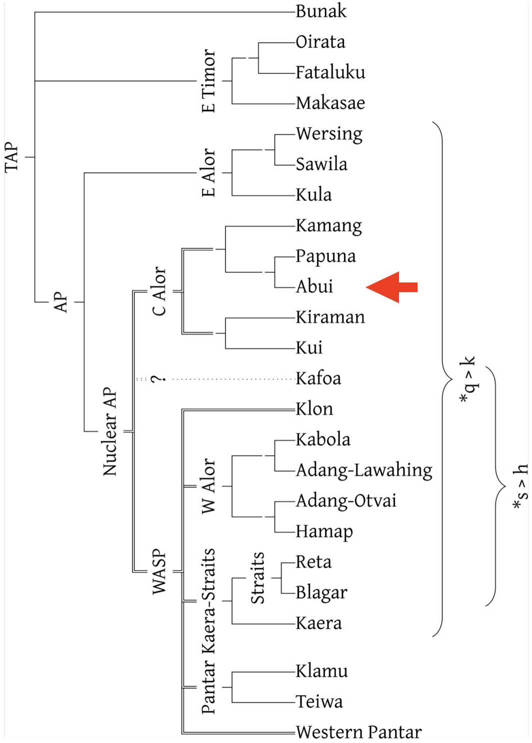


Alor-Pantar

Abui (Papuan, Timor-Alor-Pantar family)



© 2018 Owen Edwards and UBB



Kaiping, G., & Klamer, M. (2022). The dialect chain of the Timor-Alor-Pantar language family, Language Dynamics and Change.

Abui Wordnet development

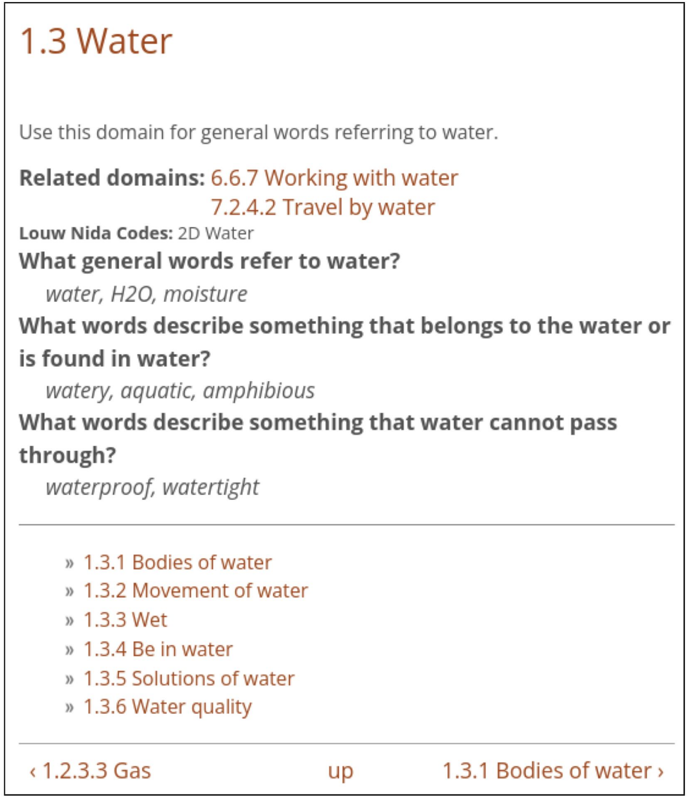
The Abui Wordnet was developed following the **expansion approach** (Kratochvíl and Morgado da Costa, 2022).

- naive multilingual sense intersection algorithm
- linking **data collected over the last two decades**
- traditional descriptive workflow (text recordings, elicitations)
- glosses in English, Indonesian and Alor Malay (regional variety)
- Abui Wordnet v1.0 contained 1,475 synsets and 3,606 senses
- **entirely hand-checked** by B. Delpada (native speaker of Abui)
- released under **CC-BY 4.0 license**
- Can be found here: https://github.com/fanacek/abuiwn

SIL Semantic Domains: Structure and Use

The SIL SemDoms (**http://semdom.org/**) is an **ontology** created by the Summer Institute of Linguistics linguist 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 as a hierarchy
- Words are grouped by topics
- Each SemDom includes questions that elicit synonyms and related senses
- Common sense knowledge



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 ›

Rapid Word Collection Workshops

The Rapid Word Collection (RWC) method **accelerates lexicographic work** by:

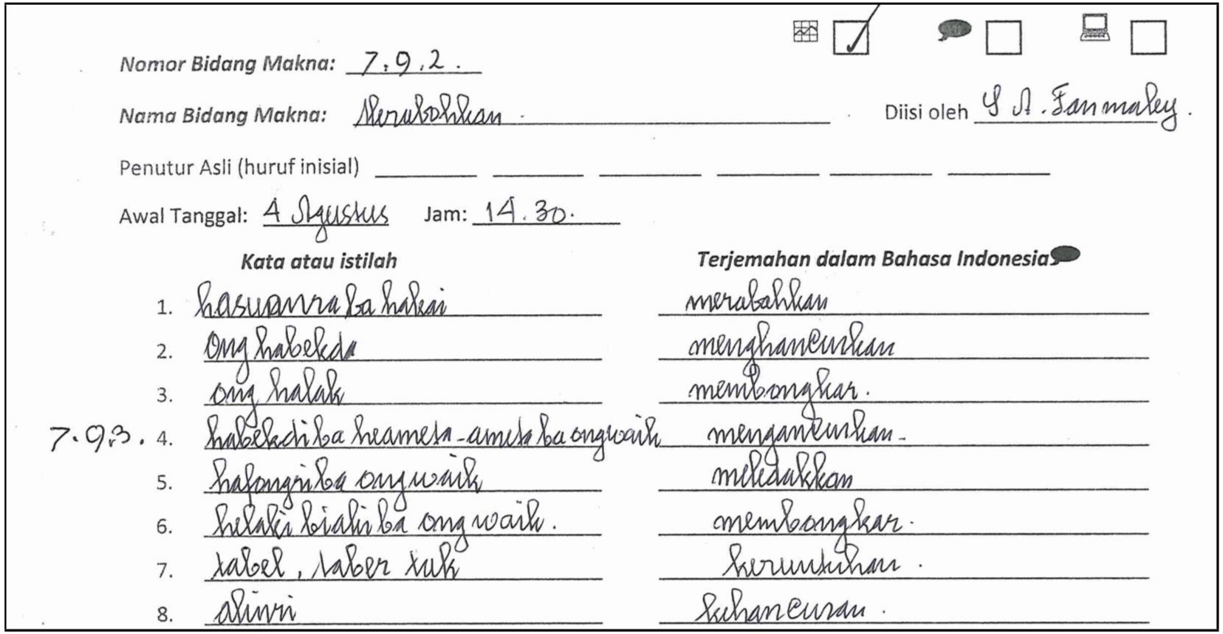
- involving language communities (can handle limited literacy)
- prompting associative memory by questions (not tiring for participants)
- distributing the work across groups (competitiveness, enthusiasm)

A two week workshop regularly yields over 10 000 entries (= many years for a linguist working alone)

- lexicon entities have multiple senses
- coverage not biased by the corpus composition
- corpora upward of a million words would yield as many entries
- community building and awareness raising potential (vitality and complexity)

Abui Rapid Word Collection Workshops

SemDoms in Indonesian; 3 RWC Workshops for Abui (in 2013, 2014, and 2016); 🌞 10 working days 👷 25 people involved on any day (in total 67 men 21 women), groups of 3 (except 2 teachers) +2000 person/hours; 📖 over 17k raw entries > 12,4k digitized > over 1000 person/hours in digitization, translation and checking



☑ □ □

Nomor Bidang Makna: 7.9.2.

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

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

Awal Tanggal: 4 Agustus Jam: 14.30.

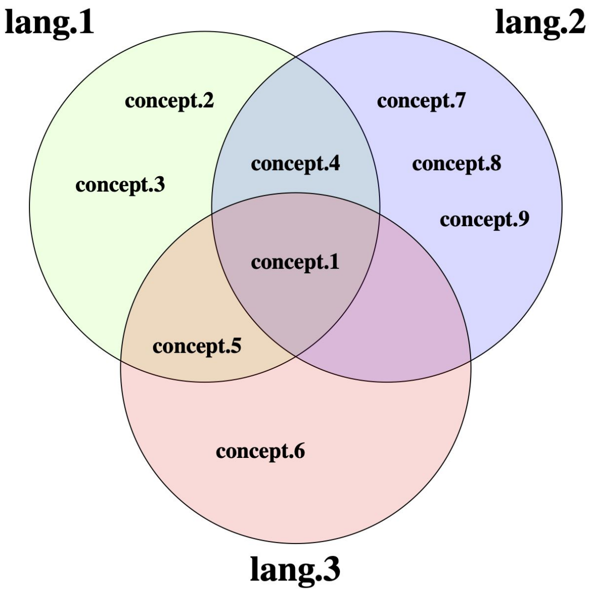
	Kata atau istilah	Terjemahan dalam Bahasa Indonesia
1.	hasuanra ba hakai	merobohkan
2.	Ong habekda	menghancurkan
3.	ong halak	membongkar.
7.9.3. 4.	habekdi ba heamesa-amesa ba ongwaik	mengancurkan.
5.	hafongri ba ongwaik	meledakkan
6.	helaki biaki ba ong waik.	membongkar.
7.	tabel, taber tuk	keruntuhan.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

Words are recorded in worksheets that track the semantic domain ID, the name of the language consultants and often provide Indonesian or Alor Malay translations.

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

Methodology

We use and extend the idea of **Multilingual Sense Intersection** (**MSI**, Bond et al., 2008; Bonansinga and Bond, 2016) – used to build the Coptic Wordnet (Slaughter et al., 2019) and to kick-start the Abui Wordnet.



We **restrict the available semantic space of words through the intersection of aligned translations** of that same word.

Step 1: apply MSI to link Abui RWC data to OMW

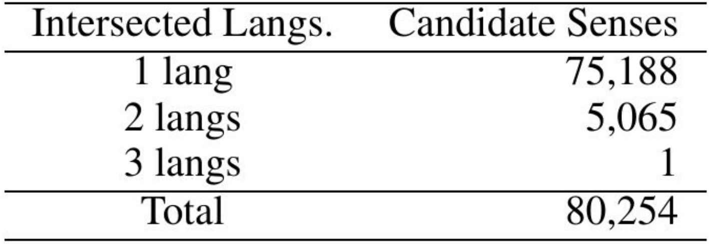
Step 2: apply MSI to link SemDoms to OMW

Step 3: intersect both mappings to find new candidate senses for AbuiWN

Linking the Abui RWC data to Wordnet

The RWC workshops generated **11,657 Abui lemmas** (i.e. digitized):

- 11,638 translated to Indonesian
- 9,078 translated to Alor Malay
- 5,846 translated to English



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

Only 1 word was intersected by 3 languages, and we know this is not enough:

- Coptic Wordnet (Slaughter, et al., 2019).:

  1-lang = 7-25%, 2-langs = 49%-89%, **3-langs = 63%-98%, 4-langs = 100%**

- Abui Wordnet (Kratochvíl and Morgado da Costa, 2022):

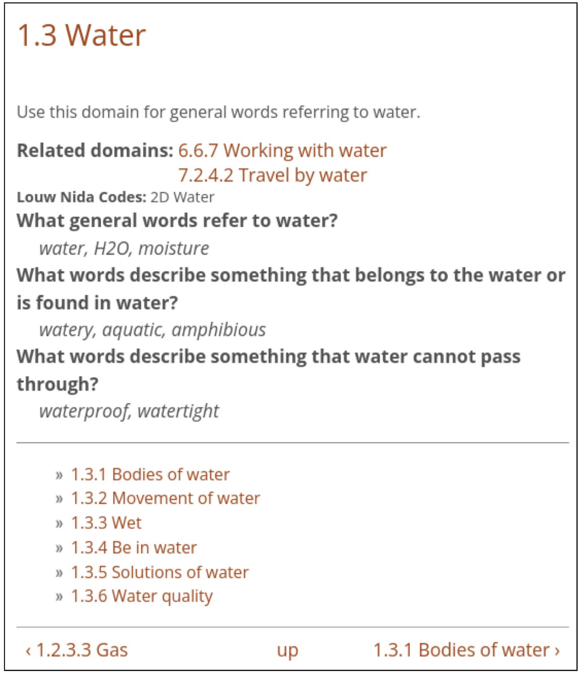
  1-way = 35%, 2-way = 50%, **3-way = 99%**

Linking the Abui RWC data to Wordnet

The RWC workshops generated **11,657 Abui lemmas**:

- 11,638 translated to Indonesian
- 9,078 translated to Alor Malay
- 5,846 translated to English
- **11,657 linked to SIL SemDoms**

However, **SIL SemDoms are only identifiers**, with many words, prompted by different questions.



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

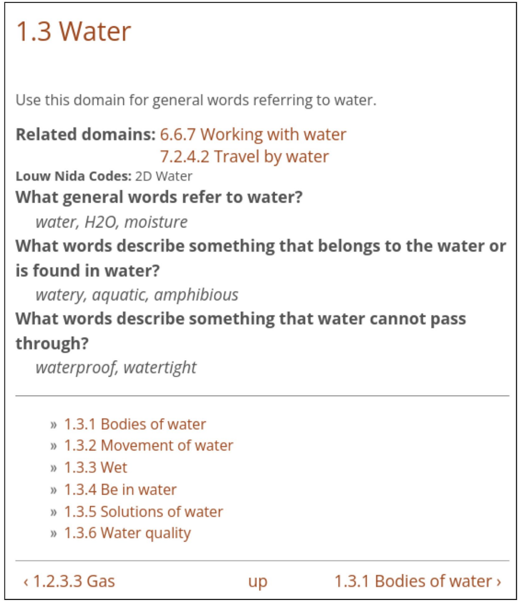
‹ 1.2.3.3 Gas up 1.3.1 Bodies of water ›

Solution [???]

- **Link all words within a SemDom to wordnet**
- Use the inventory of concepts within a SemDom to **filter sense candidates**

Linking SIL SemDoms to Open Multilingual Wordnet

- Linking SIL SemDoms to OMW **first proposed in Rosman et al. (2014)**
  - Access to only 2 languages (eng, ind)
  - Linked only SemDom titles/headers (1,792)
    - RWC aims to collect words within a SemDom
    - Tennis-problem (*tennis*, *racket*, *ball*, *net*)
- Our approach differs in two key aspects:
  - We used data for **13 languages**
  - We **linked both *titles* and *example words***



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

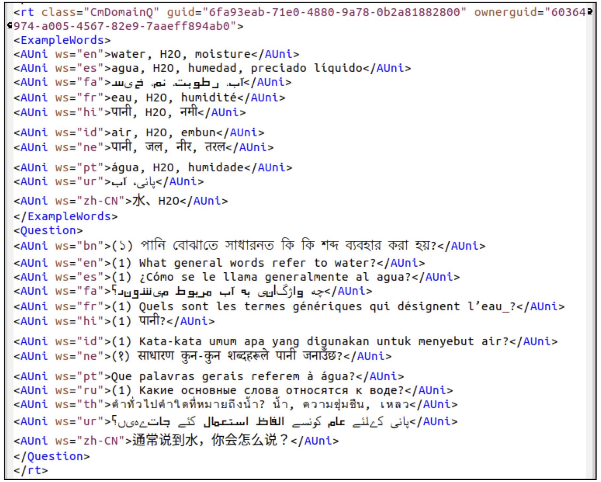
‹ 1.2.3.3 Gas up 1.3.1 Bodies of water ›

Extracting SIL SemDom Data from SIL Fieldworks

SemDoms (**CC-BY SA 4.0**) used in a few language documentation tools (e.g., SIL Toolbox, **SIL Fieldworks**, SIL Lexique Pro, WeSay) – but are **not very easy to find** (in machine readable form)!

We extracted **SIL SemDom data from SIL Fieldworks** – translations in 14 languages, provided by volunteer linguists (**unbalanced/incomplete**)

Some normalization per language was required (e.g., difference chars for spaces, punctuation)



```
<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, ความชุ่มชื่น, 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>
```

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

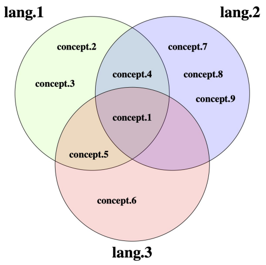
Temporarily Expanding the Open Multilingual Wordnet

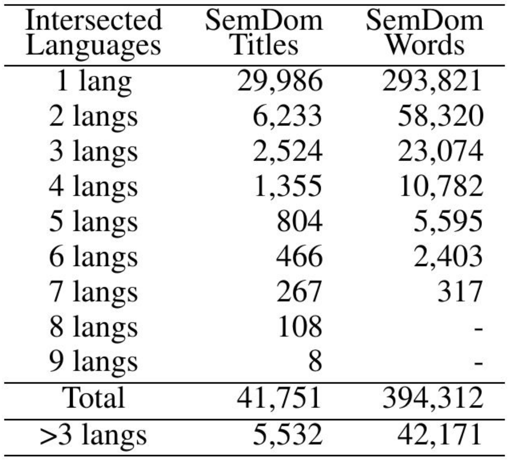
13 Languages	**Wordnet used for linking**
English	Princeton WordNet **(Fellbaum 1998)**
French	WOLF **(Sagot and Fišer 2008)**
Spanish	Multilingual Central Repository **(Gonzalez-Agirre et al., 2012)**
Chinese	Chinese Wordnet **(Huang et al. 2010)**, Chinese Open Wordnet **(Wang and Bond 2013)**
Portuguese	OpenWordnet-PT **(de Paiva and Rademaker 2012)**
Indonesian	Wordnet Bahasa **(Mohamed Noor et al. 2011)**
Farsi	Persian Wordnet **(Montazery and Faili 2010)**
Thai	Thai Wordnet **(Thoongsup et al. 2009)**
Hindi, Urdu, Nepali, Bengali	IndoWordnet **(Bhattacharyya, 2010)**, IndoWordnet-PWN Mapping **(Kanojia et al., 2018)**
Russian	Russian Wordnet **(Loukachevitch et al., 2016)**, and PWN mapping **(Loukachevitch and Gerasimova, 2019)**

Linking SIL SemDoms to Open Multilingual Wordnet

Applied **MSI for the SemDom data**:

- Separate candidates for *Titles* and *Words*
- **Largest intersections had 9 languages** (max. 13)
- 41,700 candidates from SemDom titles, and about 394,000 candidates for SemDom example words
- We assume as useful **only data with >=3 langs**
  - 5,500 SemDom titles; 42,000 SemDom words
- 1,173 out of 1,792 (65%) SemDom titles get >=1 link
  - 2.5.6 *Symptom of disease*; 5.8 *Manage a house*
- 1,671 out of 1,792 (93%) SemDoms get >=1 word link



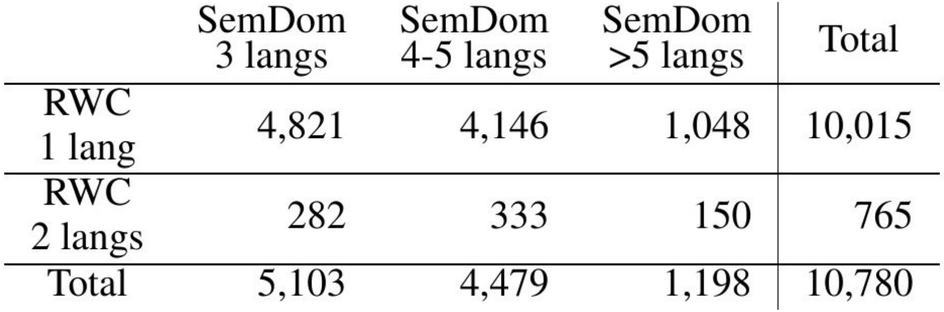


Intersected Languages	SemDom Titles	SemDom Words
1 lang	29,986	293,821
2 langs	6,233	58,320
3 langs	2,524	23,074
4 langs	1,355	10,782
5 langs	804	5,595
6 langs	466	2,403
7 langs	267	317
8 langs	108	-
9 langs	8	-
Total	41,751	394,312
>3 langs	5,532	42,171

Filtering/Intersecting RWC Data with SemDom Links

- We **filtered sense candidates** produced by RWC data with SemDom links
- Keep only about **13% of candidates** (>80,000 candidates)
- Separated this data into **6 classes** (2 axis):
  - Intersected languages through RWC (1 or 2 languages)
  - Intersected languages informing the SemDom links (3 classes)

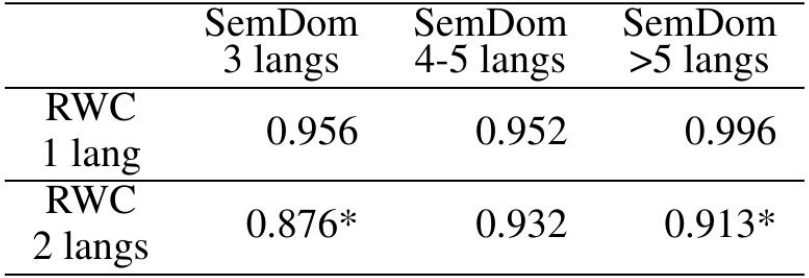
Assumption: the higher the intersection level of both axes, the higher the quality of the suggested senses



	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

Hand-Checked Evaluation

- Hand-checked **~250 random senses from each of the six classes** (1,432)*
  - 2 native speakers, 2 expert linguists
- **Directly evaluates automatic sense linking** for the Abui Wordnet and **indirectly evaluates SemDom linking**
- **Good overall results**, but surprising for [RWC 2 langs & SemDom > 5 langs]
  - Error analysis discovered quite a few incorrect senses in the Wordnet Bahasa (automatically translated senses, e.g., *draw*)



	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*

Release Notes

- New data for the **Abui Wordnet** is released under a **CC BY license**
  - Github (link in the paper)
  - Only 248 of 10,780 generated senses were already in the wordnet
  - We will finish checking the data before a new release
- The new **SIL SemDom Links** will be released under **CC-BY-SA license**
  - Github (link in the paper)
  - Two TSV files (SemDom Titles, SemDom Words)
  - All data (no filtering) ~41,700 links for SemDom titles, and ~394,000 candidates for SemDom titles

Future Work

Finish Hand-checking: 1,432 senses checked, ~9,300 candidate senses to do;

Finish Digitization: 12,421 digitized; 12,415 Indonesian translation; 9,168 Malay translation; 6,312 English translation; orthographic variation kept (~50% through);

Improve SemDom linking: Explore different language pairs; Use wordnet hierarchy; Use more data from SemDom (e.g. questions vs. definitions);

SIL Integration: Clean-up/Hand-check SemDom links to OMW, automatically populate translations into other languages; suggest more words to each question;

Lemma forms: possessive prefixation on nouns, object agreement on verbs;

Sense tagging: speed up by applying WSD on the English-Abui bitext [?]

Far future: Contribute to / incorporate the Wordnet Bahasa (semi-orphaned)

Concluding Remarks

- Used a **naive yet powerful MSI algorithm** to expand the Abui Wordnet
  - Automatically linked over 10,000 senses with >90% accuracy
- Created a **new and improved mapping of SIL Semantic Domains** to OMW
  - >430,000 links to OMW (>47,000 high quality links)
  - Very relevant to **linking data in field linguistics**
  - We need to showcase the **immediate benefits of linking data** to OMW (e.g. multilingual links: dictionary building, automatic glossing, etc.)
  - Wordnets can be used to **better serve underprivileged language communities** (e.g. language documentation, revitalization, education)

References

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

Bonansinga, Giulia 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.

Bond, Francis, 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.

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

Fellbaum, Christiane (ed). 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Gonzalez-Agirre, Aitor, 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.

Huang, Chu-Ren, 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).

Kanojia, Diptesh, 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).

Kratochvíl, František 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.

Loukachevitch, Natalia 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.

Loukachevitch, Natalia, 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.

Louw, Johannes P., 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.

References

Mohamed Noor, Nurril Hirfana, 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.

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

Morgado da Costa, Luís 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.

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

Rosman, Muhammad Zulhelmy bin Mohd, 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.

Sagot, Benoît 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.

Seah, Yu Jie 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.

Wang, Shan 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).

Slaughter, Laura, 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.

Thoongsup, Sareewan, 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.

thank you