Abui Wordnet: Using a Toolbox Dictionary to develop a wordnet for a low-resource language

František Kratochvíl and **Luís Morgado da Costa**
Department of Asian Studies
Palacký University Olomouc
Czech Republic
frantisek.kratochvil@upol.cz and luis.morgadodacosta@upol.cz

Abstract

This paper describes a procedure to link a Toolbox dictionary of a low-resource language to correct synsets, generating a new wordnet. We introduce a bootstrapping technique utilising the information in the gloss fields (English, national, and regional) to generate sense candidates using a naive algorithm based on multilingual sense intersection. We show that this technique is quite effective when glosses are available in more than one language. Our technique complements the previous work by (Rosman et al., 2014) which linked the SIL Semantic Domains to wordnet senses. Through this work we have created a small, fully hand-checked wordnet for Abui, containing over 1,400 concepts and 3,600 senses.

1 Introduction

This paper describes the development of a wordnet for Abui, one of more than twenty Timor-Alor-Pantar languages of Eastern Indonesia. The Timor-Alor-Pantar (TAP) languages are a western outlier among other Papuan languages, the bulk of which are spoken in and around the island of New Guinea. While the TAP languages constitute a coherent family (Holton et al., 2012; Kaiping and Klamer, 2022), their relationship to other Papuan families of New Guinea has not been demonstrated (Holton and Robinson, 2014; Schapper et al., 2014).

Within the TAP language family dictionaries exist for only a handful of languages, listed here in alphabetical order: Abui (Kratochvíl and Delpada, 2008), Blagar (Steinhauer and Gomang, 2016), Kamang (Schapper and Manimau, 2011), Sawila (Kratochvíl et al., 2014), Teiwa (Klamer, 2012), and Western Pantar (Holton and Koly, 2007). These dictionaries exist in printed form and have been also distributed in the speech community. For the remaining languages a number of wordlists exist: from 1930s the Holle lists (Holle et al., 1980), Stokhof lists (Stokhof, 1975), and various wordlists produced by the Indonesian Language Development and Fostering Agency (Pusat Bahasa Indonesia). All available wordlists are consolidated in the LexiRumah online database (Kaiping et al., 2022) which contains at least two hundred words per language.

None of the above listed TAP dictionaries contain more than 4,000 words although each of them took several years to create. Beyond the basic vocabulary, which is also included in the LexiRumah wordlists, the dictionary coverage is determined by the collected texts and the preferences of the compilers. As a result each dictionary inevitably contains random gaps. The lexicographic workflow in language description is slow, over-reliant on a single author or a small team; it does not produce lexicographic materials suitable for language revitalisation or natural language processing applications. There is generally little concern for "open" data and shared formats.

1.1 Lexicography of low-resource languages

Field linguists use a variety of lexicographic tools. Their main producer is the Summer Institute of Linguistics (SIL) which developed the SIL multi-dictionary format (MDF) described in Coward and Grimes (2000) and utilised it in the following tools:

- SIL Shoebox[1] (1st generation corpus management tool, parser, and dictionary builder)
- SIL Toolbox[2] (2nd generation corpus management tool, parser, and dictionary builder)
- SIL Lexique Pro[3] (2nd generation dictionary management tool)
- SIL FieldWorks[4] (3rd generation, with all previous functionalities, plus automated grammar generation)

[1] https://software.sil.org/shoebox/
[2] https://software.sil.org/toolbox/
[3] https://software.sil.org/lexiquepro/
[4] https://software.sil.org/fieldworks/

Proceedings of the first workshop on field linguistics, pages 54–63
October 16, 2022.

- SIL WeSay[5] (4th generation, a collaborative native-speaker oriented lexicographic tool)
- LanguageForge[6] (a web-based dictionary development tool sharing the data format with FieldWorks but running on any OS with a browser)

In addition, the Max Planck Institute introduced the MPI Lexus online plaform[7] which did not become mainstream. For comparative wordlists the comma-separated-value format is now mainstream and it is used in all Cross-Linguistic Linked Data[8] databases such as Dictionaria[9] or the The Austronesian Comparative Dictionary Online[10].

1.2 Desiderata

Lexicography of low-resource language requires tools with broad functionality. Firstly, the tools should support making fine-grained meaning distinctions (instead of 5 verbs glossed as 'cut' offer means to systematically distinguish them). The tools should also allow the lexicographer to monitor the coverage of various semantic fields to produce balanced resources. Finally, a dictionary should be structured in a way that enables its use in semantic typology.

Next, the tools should complement grammatical description, embedding information on phonetics, morphosyntax, usage, etc., and support semantic tagging of the corpus. The tools should support itegration of the lexicon and corpus, to draw naturalistic examples.

Another concern are the data formats which should rely on the maturing standards in the NLP. The interoperability with such standards is a prerequisite for gaining benefits from existing resources for major languages. For example, when identifying the most appropriate sense of a word in the low-resource language, equivalents in other major languages should be discoverable automatically.

Finally, the lexicographic tools should systematically support crowd-sourcing and community maintenance because it is unlikely that the number of professional linguists studying a low-resource language can ever become adequate for the task at hand.

We believe that wordnets are tools that meet the above desiderata and we will briefly characterise them in the next section.

2 Wordnets and Low-resource Languages

There are two main methods to build wordnets (Vossen, 1998). The first is known as the 'expansion approach', where the semantic hierarchy of another wordnet is used as pivot. In this approach, the required work is essentially a translation effort – conserving the structure of the pivot wordnet and translating individual nodes of the hierarchy, which can be done incrementally (i.e. usually starting by a subset of frequent concepts) but can take in principle infinitely long until all language specific senses are identified. The Princeton Wordnet (PWN, Fellbaum, 1998) is, by far, the most frequently used pivot for projects that employ the 'expansion approach'.

The second method is known as the 'merge approach'. And while this approach is perhaps more principled, in theory, it is both slow and it also requires more resources. In the 'merge approach' no pivot structure is assumed. As such, this method can ensure higher degrees of freedom while modeling the structure of the wordnet without depending on pre-assumed semantic relations. One of the immediate benefits of this approach is the ability to freely add new concepts that are not part of the pivot language – a problem many wordnet projects that followed the 'expansion' approach have struggled with. The major drawback of this approach, however, is its inability to immediately benefit from the parallel translations available from all other projects that used the same pivot.

2.1 The Collaborative Interlingual Index (CILI)

In recent years there have been two major changes to wordnets that have made wordnets more suitable to deal with low-resource languages. These are: the Collaborative Interlingual Index (**CILI**, Bond et al., 2016) and a new and improved Wordnet Lexical Markup Framework (**WN-LMF**, P. McCrae et al., 2021).

CILI has solved the linking problem: before CILI it was necessary to use one language as a pivot to link other languages. Historically, this pivot has been the Princeton WordNet (Fellbaum,

[5] https://software.sil.org/wesay/
[6] https://languageforge.org/
[7] https://www.mpi.nl/corpus/html/lexus/index.html
[8] https://clld.org/
[9] https://dictionaria.clld.org/
[10] https://acd.clld.org/

1998) – a decision embraced by the Open Multilingual Wordnet 1.0 (OMW, Bond and Foster, 2013), a projected that linked dozens of wordnets projects using English as a pivot language. Even though the choice of English as a pivot brought forth many benefits, this quickly became problematic to describe non-main-stream languages whose concept inventories often differ from English (i.e., many languages have senses for concepts that had not been described for English, making it quite difficult to streamline the development of a wordnet that did not largely overlap with English).

As an alternative, instead of chosing English as a pivot, wordnets were developed independently from English, but the downside of such approach is that wordnets can no longer be linked together. The independent construction has historically only been viable for very large projects, with a strong funded agenda, and is not really recommended for smaller projects.[11]

CILI, which was largely inspired by the Interlingual Index (ILI) developed for the EuroWordNet (Vossen, 1998), ended the need to use any specific language as pivot. CILI not only allows any language to contribute to a language-agnostic concept inventory, but also allows a language to link directly to other languages without using English as pivot, harnessing the advances in meaning description made in any linked language.

As an example we may give the Abui word *liik* 'elevated wooden platform' which can refer to a chair, table, a gazebo, wooden house floor, verandah, gallery or a stage and corresponds quite well to the Indonesian and Malay words *balai-balai* or *bale-bale*, which have no simple equivalent in English. English does not have a generic word describing an elevated wooden platform but usually lexicalises its size or purpose. In CILI the Abui and Indonesian/Malay words can be linked without the need to link to English.

There may also be words that are unique for Abui (and perhaps related languages) which have no counterparts in English or Indonesian, but may have one in one of the languages already linked to CILI. Examples of such words are the Abui *neura* 'sibling of the opposite gender' and *nemuknehi* 'sibling of the same gender'. Interestingly, English and Malay lexicalise the gender of the referent while Abui distinguishes the same-gender siblings (brother-brother, and sister-sister *nemuknehi*) from opposite gender (brother-sister = *neura*).

[11] This is discussed in greater detail more in Section 4.

2.2 WN-LMF

The second major breakthrough that now extends the utility of wordnets (for fieldwork or otherwise) is the improved and continuously expanding WN-LMF. Wordnets traditionally contained only open class words (i.e., nouns, verbs, adjectives and adverbs) – which immediately raised limitations on the use of wordnets as fuller lexicons. However, this restriction is no longer true, as can be seen by an increasing trend in expanding wordnets not only to other word classes – e.g., pronouns (Seah and Bond, 2014), exclamatives (Morgado da Costa and Bond, 2016), classifiers (Morgado da Costa et al., 2016) – but also to expand wordnet towards new depths of linguistics analysis, to include new layers of annotation that include better ways to represent regional or diachronic orthographic variation, pronunciation (incl. links to audio files), syntactic modeling, and much more. These efforts are constantly being updated on a need basis, and are summarized in a publicly released WN-LMF schema that strongly encourages different languages to encode this information in a shared format.

2.3 Open Multilingual Wordnet

The Open Multilingual Wordnet (OMW, Bond and Foster, 2013) is, perhaps, the best example of the benefits provided by the 'expansion approach'. The OMW currently links dozens of open wordnets using PWN as the pivot structure. The language alignments provided by all these parallel wordnets are extremely useful for many downstream NLP tasks, such as Machine Translation and Word Sense Disambiguation.

A recent change to the way the OMW operates was introduced with the creation of the Collaborative Interlingual Index (CILI, Bond et al., 2016) – an open, language agnostic, flat-structured index that links wordnets across languages without imposing the hierarchy of any single wordnet. Through CILI, multiple projects are now able to link to each other and to contribute directly to the set of CILI's concepts without the penalty of being frozen within an imposed structure.

Naturally, CILI was initially created using the concept set provided by the PWN (i.e. all PWN concepts have a direct link to CILI), the quickest and easiest way to link a new wordnet to CILI is still to use the expansion approach with PWN's

hierarchy as pivot – and this is what we chose to do.

The architecture linking multiple wordnets has been implemented in the Open Multilingual Wordnet (OMW) allowing the low-resource wordnets to be linked and studied so that their properties can inform the future development and design decisions. The authors of OMW make a strong point for unrestricted (u) or attribution required (a) license release (Bond and Foster, 2013).

The new (upcoming) version of the OMW will enforce the use of the WN-LMF, further encouraging the adoption of this schema among existing wordnets, and most certainly also encouraging further discussion on future needs to expand the WN-LMF to accommodate new/missing information.

2.4 Integration of low-resource languages into global wordnet

As described in the beginning of this section, wordnets constructed before the introduction of CILI had to either be developed independently of English (merge approach) or use the PWN as their pivot (expansion approach). An example of the merge-approach is the Yami wordnet whose authors attempted to incorporate elaborate and specific information on certain semantic domains, taking the Yami fish terminology as a test case (Yang et al., 2010). As other examples may serve the Vietnamese wordnet (Lam and Kalita, 2018), Mansi wordnet (Horváth et al., 2016) or the human-curated wordnet of Old-Javanese (Moeljadi and Aminullah, 2020), which has to rely on deep philological knowledge in the absence of native speakers.

The Cantonese wordnet (Sio and Costa, 2019) is an example of the extension approach. It is a high-quality human-curated resource derived from the Chinese Open Wordnet and the PWN.

The extension approach is suitable for automatic methods, as demonstrated by the Shipibo-Konibo wordnet (Maguiño-Valencia et al., 2018) which was derived from Spanish glosses extracted from a 1993 Spanish-Shipibo-Konibo dictionary. The outcome of the automatic linking was manually evaluated.

Our approach is the closest to that taken in the creation of the wordnets for Kristang (Morgado da Costa, 2020) and Coptic (Slaughter et al., 2019) to which we will refer in more detail in section 4.2.

3 Lexicographic resources for Abui

Abui (ISO 639-3: abz, abui1241) is a Timor-Alor-Pantar (TAP) language spoken by about 17 thousand speakers in an area stretching from the northern to the southern coast in Central Alor. Abui is classified by Kaiping and Klamer (2022) to the Central Alor branch of TAP. The work reported here focusses on the variety spoken in the village of Takalelang at the northern coast.

The earliest lexicographic work on Abui comes from the pen of two anthropologists who conducted their research in late 1930s in the Abui village of Atengmelang. Cora Du Bois, who published a monograph on the Abui culture (Bois and Kardiner, 1944), left behind extensive lexical and grammatical notes (part of the Cora Du Bois Personal Papers at the Tozzer Library, Harvard University, [CDBpapers]). Martha Maria Nicolspeyer appended to her PhD thesis an Abui-Dutch wordlist (Nicolspeyer, 1940) [N1940]. This work served as a base for W. A. L. Stokhof, who worked on Abui in late 1970s and 1980s and published the Du Bois wordlists and provided an Abui text with a grammatical commentary (Stokhof, 1975, 1984) [S1975].

Since 2003 the Abui language has been subject to more intensive study which resulted in a full grammatical description (Kratochvíl, 2007) and a dictionary primer (Kratochvíl and Delpada, 2008) [KD2008]. The dictionary is derived from a Toolbox corpus and contains only words which are attested in texts that were recorded during the documentation.

The dictionary was revised and expanded in its second edition (Kratochvíl and Delpada, 2014) [KD2014], available online and counting over 400 pages. It includes Abui-English, Abui-Indonesian and reverses, as well as a semantic ontology based on the SIL semantic domains (Moe, 2013).

Between 2013 and 2016, three Rapid Words workshops were conducted, during which about 17 thousand words [RW2016] were collected using a crowd-sourcing approach designed by the Summer Institute of Linguistics (Boerger and Stutzman, 2018). Currently, these words are being digitised and equiped with their English, Indonesian and Malay glosses before the method described here can be applied. Table 1 offers an overview of the available Abui lexicographic work to date including its size and estimation of the production time (in years). The works are identified by the abbreviations used above.

Author(s)	Type	Words	Years
N1940	dictionary	710	0.5
CDBpapers	wordlist	2063	2
S1975	wordlist	117	n.a.
KD2008	dictionary	1757	4
KD2014	dictionary	2389	6
RW2016	wordlist	>17k	>6

Table 1: Overview of the Abui lexical resources

4 Developing the Abui Wordnet

Building and maintaining a wordnet is extremely time-consuming, especially when this is done manually. For this reason, the large majority of wordnets are built by bootstrapping their development using one or more existing wordnets, referred to as the "pivot languages", as we discussed in section 2. In this section we discuss the methods to build the Abui wordnet.

4.1 Extracting Toolbox Data

The SIL Toolbox dictionaries are based on the Multi-dictionary format (MDF) by Coward and Grimes (2000). The format defines a broad range of fields which are marked by a generic ID starting with a backslash (eg. \lx, \ph, \ps, etc.). MDF is rich and versatile: it incorporates linguistic information (pronunciation, morphosyntactic properties, meaning, examples), cultural information, sources (e.g. books, narratives, speakers), etc. An example of a lemma can be seen in Figure 1 which contains the Abui verb *pok* 'split, burst'. The figure consists of two blocks. The left block in sans-serif case contains the data from the Abui dictionary. The right column and the shading is our own and separates the lemma fields into blocks and characterises their content.

The first field of each entry is a lemma (\lx), which is followed by its pronunciation (\ph) and part-of-speech (\ps). The meaning is captured by a gloss, reverse gloss, and definition in English (\ge, \re, \de), Indonesian (\gn, etc.), and Alor Malay (\gr, etc.) Finally, the entry also contains an example sentence and its translations in English, Indonesian and Malay.

Figure 1 shows that there is some redundancy in the MDF format. For example the information in the gloss field (\ge, \gn, \gr) is always repeated in the reversal field (\re, \rn, \rr). The definition field (\de, \dn, \dr) may occasionally contains more information than the gloss and the reversal, as it is the case also in the lemma for *pok* 'split, burst' above.

\lx pok	⇐ **lemma**
\ph ˈpok	⇐ **pronunciation**
\ps v.0 \pn kki	⇐ **part of speech**
\ge split \re broken \re smashed \re split \de split, burst, hatch, broken, smashed	⇐ **gloss, reversal, and definition (ENG)**
\gn pecah \rn retak \rn menetas \rn pecah \dn pecah, menetas, retak	⇐ **gloss, reversal, and definition (IND)**
\gr peca \rr menetas \rr peca \dr peca, menetas	⇐ **gloss, reversal, and definition (MLZ)**
\ref Poku.001 \xv Pingai nu hayei poku. \xe A plate fell down and broke. \xn Piring itu jatuh dan pecah. \xr Piring tu jatu peca.	⇐ **example sentence and translations**

Figure 1: The lemma for *pok* 'split, burst' (MDF format)

Abui Lemmas	2,508
English Lemmas	4,985
English Definitions	2,766
Indonesian Lemmas	3,829
Indonesian Definitions	5,771
Malay Lemmas	3,267
Malay Definitions	2,633

Table 2: Summary of data extracted from Toolbox

For the work presented in this paper, we used only the information contained in the Abui lemma, part-of-speech, reverse glosses (referred as individual language lemmas) and definitions. Table 2 provides a summary of the amount of information extracted from the Abui Toolbox dictionary.

The table reveals that the number of Indonesian definitions is higher than the number of lexemes because different senses of the word were included under the same lexeme, such as *aha* for which three senses were listed: (i) 'outside', (ii) 'outside, in the fields' and (iii) 'blade, the sharp part of a cutting tool'. Each sense contains a separate definition, but the reverse glosses for Indonesian are shared across all available senses.

4.2 Multilingual Sense Intersection

In our work, we exploit the existing lexicographic work on Abui to bootstrap the development of the wordnet following the expansion approach

while acquiring sense candidates through a naive algorithm inspired by multilingual sense intersection (Bonansinga and Bond, 2016; Bond and Bonansinga, 2015) to determine potential senses of a new wordnet – a similar method to the one employed to build Coptic Wordnet (Slaughter et al., 2019), while using field data instead of dictionary data.

Multilingual sense intersection has a simple logical foundation. The base idea is that the semantic space of a polysemous word in any language can be constrained by aligned translations of the same word in other languages. This same concept has been used in automatic Word Sense Disambiguation (WSD) using parallel text. And using data with an increasing number parallel languages has been shown to incrementally improve the sense disambiguation. In our case, however, instead of using parallel text to disambiguate multiple languages at the same time, we use existing wordnets as pivots to generate candidate senses for a new wordnet. Figure 2 shows a conceptualization of this logic, for three languages.

We used available wordnet data for the three languages present in our Toolbox data – English, Indonesian and Alor Malay (a Vehicular Malay variety). English wordnet data came primarily from the Princeton Wordnet (Fellbaum, 1998). Indonesian and Malay data came primarily from Wordnet Bahasa (Noor et al., 2011; Bond et al., 2014).

In addition to these wordnets, we used data made available by the Extended Open Multilingual Wordnet (Bond and Foster, 2013), which contains automatically collected data from Wiktionary and the Unicode Common Locale Data Repository (CLDR), as well as data made available through the ongoing sense annotation efforts of the NTU Multilingual Corpus (Tan and Bond, 2014; Bond et al., 2021) – which have expanded the sense inventory of the above mentioned wordnets.

Figure 2 illustrates a hypothetical scenario where a single Abui lemma is a candidate sense for nine possible concepts (*concept.1–9*). However, these nine senses are not all equally suggested by the three languages. In this example, the available English (ENG) translations suggest five concepts, the Indonesian (IND) translations also suggest five concepts (although not the same five), and the Malay (ZSM) translations suggest three concepts.

A natural way to organize this data is by the number of languages that suggest any given sense.

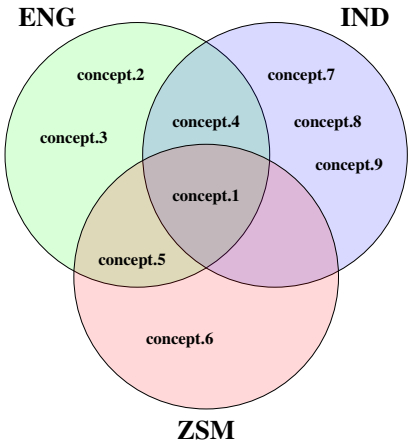

Figure 2: Sense Intersection visualisation: coloured circles represent lexemes which refer to a number of senses (concept.1-9). Unrelated languages are less likely to colexicalise the same set of senses.

In our example, *concept.1* would be suggested by all three languages, while both *concept.4* and *concept.5* would be suggested by alignments in only two languages. Empirically, it is easy to understand that senses suggested by more languages have a higher likelihood of being correct.

In addition to determining the number of intersected languages, our current algorithm also uses other simple metrics to rank Abui sense candidates, including: number of individual senses matched within a concept for each language (each worth ten points); and the number of matches between an existing wordnet sense and the definition extracted through Toolbox (each worth one point).

Since most synsets in wordnet have more than one sense, the ranking score in our algorithm seeks to reward candidates that show a greater overlap with the information contained in each wordnet. This means, for example, that the Princeton Wordnet concept for the verb `00056930-v (cause to be born)`, which has five difference senses (*bear; have; birth; deliver; give birth*), would contribute with a score of ten points for each lemma that was included in the English translations of Abui Toolbox dictionary entry for the corresponding verb. Scores gathered by each language are summed into a final score.

In order to reduce spurious candidates, only data with congruent parts-of-speech (between the wordnets and the Toolbox data) was used. This was done by creating a hand mapping between the fine-grained parts-of-speech labels included in the Toolbox dictionary, and the simpler tags that used in wordnets.

4.3 Results

The results of our sense intersection experiment are summarized in Tables 3 and 4. Table 3 shows the results in relation to the number of languages that were intersected for each sense candidate. As it would be expected, three-way intersection happens much less frequently than two-way intersection or than senses suggested by a single language. We hand-checked all 2,368 candidate senses suggested by the intersection of three languages. In addition, for candidates informed by either two or one language, we performed a stratified sampling (based on score bands shown in Table 4) and checked an extra 1,200 candidate senses. From this evaluation, we can show that senses suggested by three languages were correct around 99% (0.989) of the time, followed by 50% accuracy for senses suggested by two languages, and 35% of the time for senses suggested by a single language.

These results are in line with those reported by Slaughter et al. (2019), for the Coptic Wordnet, where senses triangulated by three languages were shown to be correct as high as 98% of the time. Our findings are also in line with other similar work, such as Bond and Ogura (2008), who found scores of about 97% when aligning lexicons with three languages.

Table 4 shows a more detailed picture of our sense intersection experiment. It shows results filtered for different language pairings (for the case of two-way intersection), and also filtered by difference score bands for the same type of intersection. The scoring method was briefly described in Section 4.2.

One interesting aspect shown in Table 4 is the fact that two-way language intersection was comparable across all language pairs. Given the proximity between Indonesian and Malay, one would expect that intersection of English with one of the two other languages would result in better sense candidates – but this was not the case. Table 4 also shows that the naive scoring algorithm that expanded the simple metric of number of intersected languages reported in Slaughter et al. (2019) is useful enough to differentiate between candidates that received the same broad triangulation type. Candidates with higher scores in the same intersection type are correct more often. These differences become increasingly relevant the fewer the languages that inform that sense candidate. For senses suggested by a single language, we can see that higher ranking scores (which reflect that more than one sense in that language was match for a single concept) can be extremely useful to discern likely candidates. In our data, the most extreme case can be seen for English, where senses presenting a ranking score of 10 (i.e., informed by a single English sense) have an average accuracy of 7% but senses with a score between 20 and 32 (informed by more than one English sense) have an average accuracy score of 67%.

Intersection	Cand.	Sample	Acc.
3 languages	2,368	2,368	0.99
2 languages	8,115	600	0.50
1 language	28,678	600	0.35

Table 3: Summary of results filtered by number of intersected languages

Intersection	Cand.	Score	Samp.	Acc.
eng+ind	3,032	31-61	100	0.61
eng+ind		20	100	0.34
eng+zsm	206	21-31	60	0.65
eng+zsm		20	140	0.44
ind+zsm	4,877	31-63	100	0.61
ind+zsm		20	100	0.42
eng	9,716	20-32	100	0.67
eng		10	100	0.07
ind	17,380	21-32	100	0.57
ind		10	100	0.11
zsm	1,582	11-21	100	0.44
zsm		10	100	0.22

Table 4: Summary of results for one and two-way intersection filtered by languages and ranking score

These results show that even though our ranking algorithm is very naive, we are moving in the right direction. It would most certainly be beneficial to improve our ranking algorithm with other classic features used in Word Sense Disambiguation, such as exploiting the semantic hierarchy or using wordnet glosses and definitions.

5 Release and Licensing

A summary of the size and part-of-speech coverage of the first release of the Abui Wordnet is given in Table 5. This first release includes only data derived from candidates generated by three-way intersection – which we showed yielded data with a confidence score of 99%. Since all candidates intersected by three languages were hand-checked, we include only those that were confirmed. In addi-

tion, compatible morphological alternations were added, semi-automatically (using Toolbox data) to each sense. This increased the number of available senses considerably.

Note the low number of adjectives (which include quantifiers) and adverbs in Table 5, which is a consequence of Abui having just a handful of adjectives and encoding other properties as stative verbs, and similarly expressing event properties mostly by finite verbs (Kratochvíl, 2007, 109-110).

POS	No. Synsets	No. Senses
nouns	818	1,466
verbs	590	2,013
adjective	46	82
adverb	21	45
Total	1,475	3,606

Table 5: Abui Wordnet Coverage (v1.0)

One key motivation for this project was to inspire other field linguistics to follow on our footsteps and release their data using open licenses. Field linguists have a responsibility towards the communities they work with, and should embrace an open-shared ownership of the work that is developed with the help of these communities.

We want to encourage other field linguistics to use and replicate our work, while working towards the maintenance and preservation of Abui and its community. For this reason, the Abui Wordnet is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[12]. We have produced OMW tsv files, which can also be used in the Python Natural Language Toolkit (Bird et al., 2009). In addition, and keeping up with the recent requirements to belong to the OMW, we will also release this data using the WN-LMF format[13].

The Abui Wordnet data will be made available on GitHub at `https://github.com/fanacek/abuiwn`.

6 Discussion and Future Work

We have sketched a procedure that facilitates the transfer of the Toolbox MDF-formatted data into a wordnet. And we have also shown that it is possible to generate very high quality data through a naive algorithm based on sense intersection.

We believe our results could be improved further by improving our sense intersection algorithm to include, for example, semantic domain information,[14] or by attempting to exploit other available information often used in the task of Word Sense Disambiguation such as wordnets' semantic hierarchy, glosses and definitions.

In addition, we would like to work towards including pronunciation, grammatical information (aspectual class, valency, etymology and borrowings) and example sentences, all of which we track in the Abui Toolbox dictionary, and which can be accommodated by the Wordnet Lexical Markup Framework (**WN-LMF**, P. McCrae et al., 2021).

In the near future we also expect to have to deal with many specific features particular to Abui: (i) concepts unique to Abui or the region; (ii) extensive specific taxonomy for animals and plants[15]; (iii) many non-lexicalised CILI concepts in Abui (especially linked to technology and modernity).

Finally, another challenge we would like to work on relates to the fact that there is no official Abui orthography and many writing conventions exist which reflect dialectal and idiolectal variation as well as individual preferences regarding the spelling of vowel length, velars and uvulars, tone, and clitics. We are taking an aggregating approach and register all examples of alternative spelling and link them to the respective lemma. In the future we would like to use the full extent of the WN-LMF to make this information available in our wordnet.

7 Conclusion

This paper shows the viability of the intersection method in rapid building of wordnets for low-resource languages using data collected in field linguistics. Applying a similar method as Slaughter et al. (2019) we have reached a overall accuracy of 99% when the sense is defined by the intersection of three languages. The accuracy however does drop steeply when fewer than three languages are available.

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

[13] https://github.com/globalwordnet/schemas

[14] Semantic domains (`http://semdom.org/`) is an ontology organised in an associative way, grouping words used to talk about an area together, regardless of the subtle differences among them. For example, the English domain Rain includes words such as *rain*, *drizzle*, *downpour*, *raindrop*, *puddle*. The ontology tracks both collocations, as well as paradigm forms such as synonyms, antonyms, generic and specific relations. For example, *fly* will contain a reference to *bird* as a prototypical agent of that event. While *bird* is a generic term *chicken* is more specific.

[15] Blake, A.L. 2018. Documenting environmental knowledge in Abui, a language of eastern Indonesia. London: SOAS University of London, Endangered Languages Archive. `https://www.elararchive.org/dk0574`.

Acknowledgements

The authors acknowledges 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).

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit NLTK*. O'Reilly Media, Inc.

Brenda H. Boerger and Verna Stutzman. 2018. Single-event rapid word collection workshops: Efficient, effective, empowering. *Language Documentation & Conservation*, 12:147–193.

Cora Alice Du Bois and Abram Kardiner. 1944. *The people of Alor; a social-psychological study of an East Indian Island*. Univ. of Minnesota Press, Minneapolis,. With analyses by Abram Kardiner and Emil Oberholzer. ill.

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 Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 56–61, Trento.

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, 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, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, (57):83–100.

Francis Bond and Kentaro Ogura. 2008. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2):127–136.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.

David F Coward and Charles E Grimes. 2000. *A guide to lexicography and the Multi-Dictionary Formatter*. SIL International, Waxhaw, North Carolina.

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

K. F Holle, W. A. L Stokhof, Lia Saleh-Bronkhorst, and Alma E Almanar. 1980. *Holle lists: vocabularies in languages of Indonesia*. Dept. of Linguistics, Research School of Pacific Studies, Australian National University, Canberra.

Gary Holton, Marian Klamer, František Kratochvíl, Laura C. Robinson, and Antoinette Schapper. 2012. The Historical Relations of the Papuan Languages of Alor and Pantar. *Oceanic Linguistics*, 51(1):86–122.

Gary Holton and Mahalalel Lamma Koly. 2007. *Kamus Pengantar Bahasa Pantar Barat: Tubbe - Mauta - Lamma*.

Gary Holton and Laura C. Robinson. 2014. The Linguistic Position of the Timor-Alor-Pantar Languages. In Marian Klamer, editor, *The Alor-Pantar Languages: History and Typology*, pages 155–198. Language Science Press, Berlin.

Csilla Horváth, Ágoston Nagy, Norbert Szilágyi, and Veronika Vincze. 2016. Where bears have the eyes of currant: Towards a Mansi WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 131–135, Bucharest, Romania. Global Wordnet Association.

Gereon A. Kaiping, Owen Edwards, and Marian Klamer. Lexirumah 3.0.0 [online]. 2022.

Gereon A. Kaiping and Marian Klamer. 2022. The dialect chain of the Timor-Alor-Pantar language family: A new analysis using systematic Bayesian phylogenetics. *Language Dynamics and Change*, 12(2):274 – 326.

Marian Klamer. 2012. *Kosa kata Bahasa Teiwa-Indonesia-Inggris (Teiwa-Indonesian-English glossary)*. Language and Culture Unit UBB, Kupang.

František Kratochvíl, Isak Bantara, and Anderias Malaikosa. 2014. Sawila-English dictionary. Ms.

František Kratochvíl and Benidiktus Delpada. 2008. *Kamus Pengantar Bahasa Abui: Abui – Indonesian – English Dictionary*. UBB-GMIT, Kupang, Indonesia.

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

František Kratochvíl. 2007. *A grammar of Abui: a Papuan language of Alor*. LOT, Utrecht.

Khang Nhut Lam and Jugal Kalita. 2018. Constructing vietnamese wordnet: A case study. In *19th International Conference on Computational Linguistics and Intelligent Text Processing, March 18 to 24, 2018, Hanoi, Vietnam*.

Diego Maguiño-Valencia, Arturo Oncevay-Marcos, and Marco A. Sobrevilla Cabezudo. 2018. WordNetshp: Towards the building of a lexical database for a Peruvian minority language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ron Moe. 2013. Semantic domains. `http://semdom.org`. (Accessed 2022-08-01).

David Moeljadi and Zakariya Pamuji Aminullah. 2020. Building the old Javanese Wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2940–2946, Marseille, France. European Language Resources Association.

Luis Morgado da Costa. 2020. Pinchah Kristang: A dictionary of kristang. In *Proceedings of the Globalex2020 at the 12th Edition of the Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association (ELRA).

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.

Luís Morgado da Costa, Francis Bond, and Helena Gao. 2016. Mapping and generating classifiers using an open chinese ontology. In *Proceedings of the 8th Global WordNet Conference (GWC 2016)*, Bucharest, Romania.

Martha Margaretha Nicolspeyer. 1940. *De sociale structuur van een Aloreesche bevolkingsgroep*. V. A. Kramers, Rijswijk (Z.-H.).

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.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The GlobalWordNet formats: Updates for 2020. In *Proceedings of the 11th Global WordNet Conference (GWC 2021)*, Pretoria, South Africa. Global Wordnet Association.

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.

Antoinette Schapper, Juliette Huber, and Aone van Engelenhoven. 2014. The relatedness of Timor-Kisar and Alor-Pantar languages: A preliminary demonstration. In Marian Klamer, editor, *Alor-Pantar languages: History and typology*, pages 99–154. Language Science Press, Berlin.

Antoinette Schapper and Marten Manimau. 2011. *Kamus Pengantar Bahasa Kamang-Indonesia-Inggris: Introductory Kamang-Indonesian-English Dictionary*. Unit Bahasa Dan Budaya, Kupang.

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.

Joanna Ut-Seong Sio and Luis Morgado Da Costa. 2019. Building the Cantonese Wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 206–215, Wroclaw, Poland. Global Wordnet Association.

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.

Hein Steinhauer and Hendrik D. R. Gomang. 2016. *Kamus Blagar-Indonesia-Inggris / Blagar-Indonesian-English Dictionary*. Yayasan Pustaka Obor Indonesia, Jakarta.

W. A. L Stokhof. 1975. *Preliminary notes on the Alor and Pantar languages (East Indonesia)*. Pacific linguistics. Series B. Dept. of Linguistics, Research School of Pacific Studies, Australian National University, Canberra. By W. A. L. Stokhof. maps ; 26 cm.

W. A. L Stokhof. 1984. Annotations to a text in the Abui language (Alor). *Bijdragen tot de taal-, land- en volkenkunde*, 140(1):106–162.

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.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Meng-Chien Yang, D Victoria Rau, and Ann Hui-Huan Chang. 2010. A proposed model for constructing a Yami wordnet. In *2010 International Conference on Asian Language Processing*, pages 289–292. IEEE.