Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2940–2946
Marseille, 11–16 May 2020

Building the Old Javanese Wordnet

David Moeljadi, Zakariya Pamuji Aminullah
Palacký University, Gadjah Mada University
Olomouc, Czechia; Yogyakarta, Indonesia
{davidmoeljadi, zakariya.aminullah}@gmail.com

Abstract

This paper discusses the construction and the ongoing development of the Old Javanese Wordnet. The words were extracted from the digitized version of the Old Javanese–English Dictionary (Zoetmulder, 1982). The wordnet is built using the 'expansion' approach (Vossen, 1998), leveraging on the Princeton Wordnet's core synsets and semantic hierarchy, as well as scientific names. The main goal of our project was to produce a high quality, human-curated resource. As of December 2019, the Old Javanese Wordnet contains 2,054 concepts or synsets and 5,911 senses. It is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0). We are still developing it and adding more synsets and senses. We believe that the lexical data made available by this wordnet will be useful for a variety of future uses such as the development of Modern Javanese Wordnet and many language processing tasks and linguistic research on Javanese.

Keywords: Old Javanese, Kawi, wordnet, less resourced language

1. Introduction

This paper discusses the process of constructing a new wordnet for Old Javanese, a language written particularly in Java Island in Indonesia between 800 AD to 1500 AD. It belongs to the Western Malayo-Polynesian branch of the Austronesian language family. Despite its importance for historical and comparative linguistics, as well as ancient history, Old Javanese remains comparatively low in digital resources. To the best of our knowledge, there are some digital resources for Old Javanese such as the online Old Javanese–English Dictionary, which is a part of SEAlang projects, and Kawi Lexicon (Wojowasito and Mills, 1980). However, there is no wordnet for Old Javanese (as well as Modern Javanese). This absence of an open-source Old Javanese wordnet fed our motivations to build it.

We aim to build a machine readable resources for Old Javanese: providing a wordnet for the language, which will also be the first wordnet for the Javanese branch of Western Malayo-Polynesian. We would like our Old Javanese Wordnet to support many Natural Language Processing (NLP) tasks, such as machine translation and grammar engineering; at the same time, to support the study of linguistics, such as historical linguistics, lexical semantics, and verb subcategorization. In addition, we would like to leverage it to build a wordnet for Modern Javanese.

2. Old Javanese

Old Javanese (or Kawi, ISO 639-2: `kaw`) is the first stage of the Javanese language which has been recorded in writing for more than 600 years. The oldest inscription written in Old Javanese is dated 25 March 804 AD and Old Javanese texts and traditions continue to flourish until the present in religious and social practice, as well as in artistic and ritual contexts (Creese, 2001). It is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Javanese branch (Eberhard et al., 2019).

The term Old Javanese is given under the agreement that it is the earliest written form of Javanese literature. It was the vehicle of the culture, politics, and religions of ancient Javanese civilization and written in charters and other inscriptional documents, treatises on religious doctrine and ritual, on ascetic and mystical practice, and ethical precepts regulating man's conduct both as an individual and as a member of ancient Javanese society, and treatises on law and lexicography (Zoetmulder, 1974).

There is no standardization in writing, thus one lexical item may have more than one written form. It must be stressed that the Old Javanese data was collected from written sources mentioned above, thus we do not have any actual recording of how it was spoken in the past. Old Javanese has been written in several varieties of the Indonesian branch of Indic 'Brāhmī'-derived scripts, most abundantly in Balinese script on palm-leaf manuscripts (Zoetmulder, 1974). There is generally no word division in Old Javanese writing (scripto continua), though scholarly text editions spell Old Javanese with spaces between words.

Old Javanese borrowed a considerable number of words from Sanskrit. Out of more than 25,500 entries in the Old Javanese–English Dictionary (Zoetmulder, 1982), more than 10,000 of them were originated from Sanskrit. However, it must be noted that Sanskrit was merely adapted into Old Javanese at the lexical and textual materials in which the words were formed by derivations according to the Old Javanese morphology, so there was no form of declination nor conjugation, as Sanskrit has (Gonda, 1952). It can be briefly said that Old Javanese was 'enriched' through an infusion of lexemes drawn from Sanskrit (Hunter, 2009), as a form of participation referred to as "Sanskrit ecumene" or "Sanskrit cosmopolis" in Pollock (1996). These terms imply that Sanskrit in the further expansion has been adopted for political, literal, and economical importance in the areas attached by Hinduism and also Buddhism.

Old Javanese is an agglutinative language with various affixes (prefixes, infixes, suffixes, and circumfixes). Its base-words with affixation may undergo changes in sounds (nasalization) and in junction (*sandhi*). For example, *uttama* "excellent" from Sanskrit, by adding the Old Javanese

circumfix *ka-...-an*, will form the abstract noun *kottaman* "excellence" (Zoetmulder, 1974). Syntactically, Old Javanese is a Verb-initial language while Modern Javanese has Subject-Verb-Object (SVO) structure.

There are several sources written in Old Javanese in the form of inscriptions and manuscripts. Some of the institutions that succeeded in digitizing the original manuscripts in open-source are namely Leiden University Library and National Library of Republic of Indonesia, but not all of them are accessible in their digital version. On the other hand, École Française d'Extrême-Orient (EFEO) has also succeeded in digitizing many edited texts, in the sense of making them searchable. Certainly, the open access to digitized records of Old Javanese would greatly benefit the academic community.

To the present day, there is no general consensus for the romanized transliteration of Indian-type scriptures with which the Old Javanese was written (Acri, 2017). There are at least two romanization systems that have been proposed, widely used, and applied in several editions of the Old Javanese texts, including the proposal of Zoetmulder (1982) in the Old Javanese–English Dictionary and the one of Acri and Griffiths (2014) which is now modified in Dániel and Griffiths (2019). The first was applied in the edition of *Arjunawiwāha* "Arjuna's Marriage" by Robson (2008) and of *Sumanasāntaka* "Death by Sumanasa Flower" by Worsley et al. (2013) with the exception of the use of *ng* for the velar nasal instead of *ŋ* (*n* with palatal hook). As for the second, we can cite, for example, the edition of *Dharma Pātañjala* "Sacred Teaching of Patañjali" by Acri (2017), of *Bhīma Swarga* "Bhīma Goes to the Place of Gods" by Gunawan (2016), and of *Candrakiraṇa* "Rays of Moon" by Aminullah (2019). Both of those romanization systems are essentially adaptations of the IAST (International Alphabet of Sanskrit Transliteration) and ISO 15919 system with some additions and changes. Table 1 lists the differences of the two systems.

Zoetmulder (1982)	ĕ	ö	ṛ/rĕ	lĕ	w	ŋ
Acri and Griffiths (2014)	ə	ə̄	r̥/rə	l̥/lə	v	ṅ

Table 1: Two romanization systems for Old Javanese

Both IAST and ISO 15919 do not have a Roman character to represent the phoneme /ə/ or schwa (known in Indonesian studies as *pepet*) that Javanese language has. Thus, Zoetmulder uses *ĕ* for short *pepet* and *ö* for long *pepet*, while Acri and Griffiths propose *ə* and *ə̄* instead. The vowels *ṛ*/*r̥* and *ḷ*/*l̥* are maintained for the words of Sanskrit origin, whereas for the Old Javanese words, the conversion will be *rĕ* or *rə* and *lĕ* or *lə*. In addition to their convention, Acri and Griffiths (2014) proposed the international standard repertoire of signs, i.e. the raised circle (°) which precedes 'independent vowels' (vowels which form a separate *akṣara*) and the median dot (·) which represents *virāma* (known in Indonesian studies as *paten*).

In building the Old Javanese Wordnet we apply the romanization system that is used in the Old Javanese–English Dictionary (Zoetmulder, 1982) with reference to the respect we owe to Zoetmulder's choices and to the fact that his system is well known to most users.[1]

3. Old Javanese–English Dictionary

The Old Javanese–English Dictionary (OJED; Zoetmulder (1982)) was compiled by a Dutch scholar, Dr. Petrus Josephus Zoetmulder, with the collaboration of Dr. Stuart Owen Robson. It is the most comprehensive and authoritative dictionary for Old Javanese. The Old Javanese Wordnet we are building is based on this dictionary. The OJED contains more than 25,500 headword entries, more than 18,000 sub-entries, nearly 8,500 indications of Sanskrit origin, and over 105,000 corpus citations from more than 120 identified sources.

The headword entries are the base-words, arranged in Latin alphabetical order. Words derived (by affixation and reduplication) are listed as sub-entries under the base-words. Most of the entries and sub-entries have meanings/definitions and corpus citations or examples in corpus. The meaning/definition field may contain a question-mark which was added by the dictionary's author to show some doubt about the correctness of his interpretation.

The realization of the idea of digitizing the OJED and making it online was authorized in 2008 by the Koninklijk Instituut voor Taal-, Land- en Volkenkunde (KITLV). The project was supported by EFEO. The staff of the Sanskrit and Tamil Publishing Service (SPS) keyed the complete text of the OJED. The SEAlang Library processed the data for online publication and hosts the OJED on its website.[2]

In the online version of the OJED, almost every entry has a page number and entry number that can serve as ID number. The character *ŋ* in the original OJED was replaced with *ṅ* in the online version to simplify display and cut and paste for users. When entering search queries, users can use the Harvard-Kyoto variant given in Table 2 which will be automatically converted into Unicode, as in the Online OJED row.

This online OJED data was employed to build the Old Javanese Wordnet. Its raw data is being cleaned and annotated in order to build a database. The information in each dictionary entry is separated or classified into headwords, homonym numbers, variants, ID numbers, definitions, etymological information, notes, synonyms, equivalents in other related languages (e.g. Malay, Modern Javanese, Balinese), scientific names, references, example sources, and examples (see Figure 1). The cleaning-up and annotation process is similar to the one described in Moeljadi et al. (2017) on building the Indonesian dictionary (Kamus Besar Bahasa Indonesia or KBBI) database. A full description of this process is outside the scope of this paper.

4. Methodology

There are two main methods to build wordnets (Vossen, 1998): the 'expansion' approach and the 'merge' approach. In the 'expansion' approach, the structure of another wordnet is used as 'pivot', conserving the structure of the pivot wordnet and translating nodes of the hierarchy. The Princeton Wordnet (PWN; Fellbaum (1998)) is, by far, the most

[1] However, the character *ṅ* is preferred instead of *ŋ*. Section 3 explains the reason.

[2] `http://sealang.net/ojed/index.htm`

Online OJED	ā	ḍ	ĕ	ö	ī	ṅ	ñ	ṇ	ṛ	ś	ṣ	ṭ	ū
Harvard-Kyoto	A	D	E	O	I	G	J	N	R	z	S	T	U

Table 2: Romanized Old Javanese characters in the Online OJED and Harvard-Kyoto

kanigara 791:2 (Skt karṇikāra) a part. kind of tree with yellow flowers, Pterospermum aceri#folium

<hw>kanigara</hw> <id>791:2</id> <etim lang="Skt">karṇikāra</etim> <def>a part. kind of tree with yellow flowers</def> <sn>Pterospermum acerifolium</sn>

Figure 1: Example of an entry in OJED, before (above) and after cleaning-up and annotation (below)

frequently used 'pivot'. In the 'merge' approach, no pivot structure is assumed. It ensures a higher degree of freedom to model the structure of the wordnet based on the language in question more carefully, without depending on pre-assumed semantic relations. This approach enables the addition of new concepts that are not part of the 'pivot' language, a problem many wordnet projects that followed the 'expansion' approach have struggled with. However, it does not benefit from the parallel translations available from all other projects that used the same pivot. Our Old Javanese Wordnet uses the 'expansion' approach with the PWN and scientific names as pivot. The Sanskrit Wordnet (Kulkarni et al., 2010) could be used as a pivot since there are many Sanskrit borrowings in Old Javanese. However, it is neither connected to the PWN synset IDs nor available in the Open Multilingual Wordnet (OMW; Bond and Foster (2013)). To the best of our knowledge, there is no open-source Sanskrit Wordnet that is connected to the PWN.

Lemmas, ID numbers, definitions, and scientific names from the cleaned-up and annotated OJED data were extracted. Afterwards, the data was split into two: items whose meaning is defined through scientific names and all other items. 173 headword entries with scientific names were extracted. Using Python, the scientific names were cross-checked with the ones in the PWN. Thirty-six of them could be matched. Table 3 shows three of them.

PWN synset	**OJED ID**	**Lemma**	**Scientific name**
12200905-n	810:9	*karṇikāra*	Pterospermum acerifolium
12761123-n	1374:6	*poh*	Mangifera
11755694-n	1356:7	*pilaṅ*	Acacia

Table 3: Scientific names as pivot in the creation of the Old Javanese Wordnet

Regarding the entries without scientific names, those having English words or translations as definition, for example "seven days, a week", "world, the earth", and "support, substratum; vessel, receptacle", were extracted. Those with question marks (see Section 3) or having a long phrase or description in the definition, e.g. "a part. kind of tree with yellow flowers", were not extracted. In addition, some language specific string normalization such as removal of the infinitival 'to' and removal of determiners preceding nouns such as 'a' or 'the' were done in order to increase matches with the PWN; thus, definitions such as "world, the earth" became "world, earth" after normalization. This does not change the meaning because the PWN has both the forms with and without 'the' having the same meaning. Using Python and Natural Language Toolkit (NLTK), the extracted and normalized definitions were cross-checked with the lemmas that belong to the 4,960 'core' concepts in the PWN (Boyd-Graber et al., 2006), a usual measure for coverage of wordnet resources.

This generated a list of more than 15,000 candidate senses, spanning over 5,000 synsets.[3] Since we had no urgency for a high coverage lower quality wordnet, we decided to build a high quality one with human validation. We expect not to cover a rich hierarchy of terms because of the limited number of contexts in ancient languages such as Old Javanese.

For our human validation task, a spreadsheet containing the PWN synset ID, OJED lemma ID, OJED lemmas, PWN English lemmas, human validations, as well as PWN English definitions and examples, was created. An example of this spreadsheet is shown in Table 4. The columns OJED lemma ID and PWN examples are not shown in the table to save some space. This method of building a wordnet with human validation is similar to the one in building the Cantonese Wordnet (Sio and Morgado da Costa, 2019). As of December 2019, the first author and the second author have been checking, correcting, and adding more data to the spreadsheet manually. The first author is a linguist who has joined three intensive summer courses in Old Javanese in 2014, 2018, and 2019. He has knowledge on wordnet and has experience annotating a corpus with wordnet synsets. The second author is a lecturer in Old Javanese at Gadjah Mada University in Yogyakarta, Indonesia.

The human validation tasks comprised:

- asserting if the candidate sense in each line provided was a correct Old Javanese sense; correct senses would be marked as ○ and incorrect senses would be marked as × (see columns V1 and V2 in Table 4)
- discussing if the validation results are different and deciding the best possible judgement. For example, *aṅisuhi*, *umisuhi*, *inisuhan* mean "to wash" (transitive) in the OJED; the first author was not sure if they mean "to cleanse one's body" and thus gave the judgement ×; on the other hand, the second author gave the judgement ○; after discussion, it was decided that this should be ○ because there is an example in the

[3]Synset is a synonym set, i.e. a set of words that are synonymous or interchangeable in some context without changing the truth value of the preposition in which they are embedded. Sense is a meaning of a word in Wordnet. Each sense of a word is in a different synset.

Synset	OJED lemma	PWN lemma	V1	V2	V	PWN definition	Note
00036362-v	*aṅisuhi*, *umisuhi*, *inisuhan*	wash	x	o	o	cleanse (one's body) with soap and water	Awj 40.8 *tĕke wĕtis inisuhan iṅ kakâṅaměr* "that calf was washed by his loving elder brother"
15167027-n	*nīla*	dark	x	x	x	the time after sunset and before sunrise while it is dark outside	-
13983515-n	*nīla*	dark	x	x	x	absence of light or illumination	-
00409440-a	*nīla*	dark	o	o	o	(used of color) having a dark hue	added
15167027-n	*pĕtĕṅ*	dark	x	x	x	the time after sunset and before sunrise while it is dark outside	-
13983515-n	*pĕtĕṅ*	dark	o	o	o	absence of light or illumination	-

Table 4: Human validation

corpus from *kakavin Arjunawijaya* "The Victory of Kārtavīrya Arjuna" (Awj) that supports the PWN definition (see columns V and Note in row 1 in Table 4)

- adding the correct PWN synset if all core synsets in the PWN do not match the sense in the OJED. For example, the OJED lemma *nīla* "dark (used of color)" does not have equivalents with any PWN core synsets; thus, the correct PWN synset was manually added. In this case, we wrote "added" in the column Note (see row 4 in Table 4).

As of December 2019, 10,154 of the total set of candidate senses in the OJED entries have been hand-checked. Out of the total number of candidate senses checked, 378 had different validation results and were discussed. Out of 378 senses that were discussed, 145 were kept. In addition, more than 500 new senses were added. Section 6 describes the results in detail.

5. Issues

This section describes some issues we found and regarded as important when we built the Old Javanese Wordnet.

5.1. Variants and derived words

As noted in Section 2, one lexical item may have more than one written form in Old Javanese because of the absence of standardization in writing. The OJED lists all of these written forms, separated by commas. During the creation of the Old Javanese Wordnet, we separated these into two parts: the main lexical item and its variants, as shown in Table 5.

If the lexical items are written without brackets in the OJED, e.g. *pariwṛta*, *pariwṛtta*, *pariwarta* "train, suite, assistant, companion", the first one is regarded as the main and the others as variants. If there are brackets, such as *(m)ahĕniṅ* "clear", we made two lexical items: one with the character(s) between brackets, i.e. *mahĕniṅ*, and the other one without, i.e. *ahĕniṅ*. Both of these are from the base (or root) *hĕniṅ* "clearness", having an intransitive prefix *(m)a-* (some sources have *m*, some don't). We have adopted the one without *m* as the main one, i.e. *ahĕniṅ*, because these forms are more frequently used in the sources than the ones with *m*, such as *mahĕniṅ*. Regarding verbs, both the active or agent-oriented form and the passive or patient-oriented form appear in the OJED in the same line separated by commas, e.g. *tumutupi*, *tinutupan*, *katutupan* "to close, to cover". The root of these forms is *tutup* "cover". The infix *-um-* lends active voice or agent-oriented dynamism and the suffix *-i* lends transitivity. The infix *-in-* or the prefix *ka-* changes it into a passive or patient-oriented verb, i.e. *tinutupan* "to be closed" and *katutupan* "to be closed".[4] In this case, only the active voice form is regarded as the main one. The passive voice forms are not regarded as variants in the Old Javanese Wordnet. However, lexical items with *ka-* which are verbal adjectives, such as *kasuwur* "scattered", are kept in the Old Javanese Wordnet. They are regarded as adjectives, not passive verbs because in the OJED they appear as separate sub-entries.

5.2. Concepts not-yet in Wordnet

During the validation process, we found lexical items in the OJED that do not match any concepts in the PWN. Some of these are illustrated in Table 6.

Old Javanese has several words that denote "great-great grandchild". One of them is *pituṅ*, which also means "great-great grandfather". The lexicalized form in Old Javanese denoting the deepest level in the kinship system is *waryaṅ* "great-great-great-grandchild; great-great-great-grandfather". The PWN does not have these concepts. In addition, many Old Javanese written sources are about religious doctrine and ritual on ascetic and mystical practice and thus, we found many words related to Hinduism and Buddhism in the OJED, such as *kamokṣan* "final liberation, release" (in the sense of union between self and God). The PWN has concepts having lexical items "liberation" and "release" but none of them are equivalent to the meaning in Old Javanese. As noted in Section 4, we have extracted 173 OJED entries with scientific names. However, only 36 of them have the equivalents in the PWN. More than 130 entries related to plants and animals (most of them are particularly found in Java or Indonesia) such as *katiṅgulun* "a part. kind of wild tree with edible fruits (Protium javanicum)" are not yet in the PWN.[5] Javanese cultural artifacts such as *reyoṅ* "a musical instrument (two kettles connected

[4] The suffix *-i* in *tumutupi* becomes *-an* in passive voice.

[5] It may be the case that the scientific names or Linnaean terms in the OJED are out of date. We will investigate this for our future work.

Lexical items in the OJED	POS	Meaning	Old Javanese Wordnet	
			Main	Variants
pariwṛta, *pariwṛtta*, *pariwarta*	noun	assistant	*pariwṛta*	*pariwṛtta*, *pariwarta*
(m)ahěniṅ, *ahniṅ*	adj.	clear	*ahěniṅ*	*mahěniṅ*, *ahniṅ*
tumutupi, *tinutupan*, *katutupan*	verb	close	*tumutupi*	(none)

Table 5: Main lexical items and variants in the Old Javanese Wordnet

Semantic field	Lexical item	Meaning in the OJED
Person	*pituṅ*	great-great-grandchild; great-great-grandfather
	waryaṅ	great-great-great-grandchild; great-great-great-grandfather
State	*kamokṣan*	final liberation, release
Time	*caturyuga*	the four ages of the world
	śaka	Saka-year
Plant	*katiṅgulun*	a part. kind of wild tree with edible fruits (Protium javanicum)
	arjuna	a part. kind of tree, Terminalis arjuna
Artifact	*reyoṅ*	a part. kind of musical instrument (two kettles connected by a bar, similar to *bonaṅ*)

Table 6: Examples of specific concepts in Old Javanese

POS	Scientific names		Core synsets		Manually added	
	Syn.	Sse.	Syn.	Sse.	Syn.	Sse.
noun	34	41	948	3,112	303	548
verb	0	0	308	835	52	78
adj.	0	0	257	1,045	140	232
adv.	0	0	0	0	12	20
Total	34	41	1,513	4,992	507	878

Table 7: The number of synsets (Syn.) and senses (Sse.) from three different sources

by a bar)" do not appear in the PWN either. We are planning to add these into the Old Javanese Wordnet.

5.3. Non-ASCII characters and search queries

One important thing to be considered when using the Old Javanese Wordnet in a website with a function for search queries is the non-ASCII characters in the romanization. The Online OJED uses the Harvard-Kyoto variant, as explained in Section 3 and illustrated in Table 2. Some websites such as Sastra Jawa[6] and Dictionnaire Héritage du Sanscrit[7] give all possible lexical items as output when the users input the characters without diacritics and allow them to choose which word they are looking up. For example, lexical items *suta* "son", *sutā* "daughter", and *sūta* "charioteer" will be given as output when *suta* is entered.

6. Statistics and results

Table 7 summarizes the number of synsets and senses in the Old Javanese Wordnet with their parts-of-speech (POS), from three sources: using scientific names as pivot, using core synsets in PWN as pivot, and manual addition.

[6] https://www.sastra.org/leksikon

[7] https://sanskrit.inria.fr/DICO/index.fr.html#stemmer

POS	No. synsets	(%)	No. senses	(%)
noun	1,285	62.6	3,701	62.6
verb	360	17.5	913	15.4
adjective	397	19.3	1,277	21.6
adverb	12	0.6	20	0.3
Total	2,054		5,911	

Table 8: Old Javanese Wordnet statistics

Table 8 provides a summary of the current state of the Old Javanese wordnet. There are slightly more synsets having an adjective part-of-speech than the ones with a verb part-of-speech. There are only twelve adverbial synsets. On average, there are 3.2 senses per adjectival synset, 2.9 senses per nominal synset, 2.5 senses per verbal synset, and 1.7 senses per adverbial synset.

In total, the first version of the Old Javanese Wordnet covers a bit over 2,000 concepts using over 5,900 senses. Many of the new wordnets are closer to around 2,000 concepts (Francis Bond, personal communication). We are aiming for around 5,000 synsets and our work is still in progress. As of December 2019, our wordnet covers 30.5% of the core PWN concepts.

7. Release

The Old Javanese Wordnet contains a license file, a readme file, and a file in tab-separated-value format used by the original OMW specifications, named `wn-kaw.tab`. The format of the TAB file is shown in Figure 2. The first column contains the synset IDs, the second one contains the lemmas, and the third one contains the variants. This corresponds to the one in Table 5.

The Old Javanese Wordnet has been released on GitHub[8] under a Creative Commons Attribution 4.0 International

[8] https://github.com/davidmoeljadi/OJW

Synset ID	Lemma	Variants
09815790-n	pariwṛta	pariwṛtta, pariwarta
00460735-a	ahĕniṅ	mahĕniṅ, ahniṅ
01332730-v	tumutupi	

Figure 2: Example of the Old Javanese TAB file

License (CC BY 4.0)[9] in order to make it fully accessible to all potential users. Keeping up with the recent changes and requirements of the Open Multilingual Wordnet (OMW), the Old Javanese Wordnet will be converted to the recent WN-LMF format,[10] developed and maintained by the Global WordNet Association.[11] The use of WN-LMF is not only required by the most recent version of the OMW, but is also an essential vehicle to access the new Collaborative Interlingual Index (CILI; Bond et al. (2016)), a single shared repository of concepts. Once linked to CILI, the Old Javanese Wordnet will be able to contribute with new concepts, present only in Old Javanese such as the ones explained in Section 5.2 and illustrated in Table 6.

8. Conclusion and future work

This paper presented the ongoing efforts to build an open-source wordnet for Old Javanese. We have motivated this project with the lack of digital resources available for Old Javanese. We have introduced our methodology, which is to use the Old Javanese–English Dictionary data and the PWN to project Old Javanese candidate senses. As of December 2019, the Old Javanese Wordnet includes over 2,000 concepts and over 5,900 senses. We have discussed some specific challenges encountered while building the wordnet and how we addressed them. We hope that this new open resource will promote a variety of future uses, including language processing tasks and linguistic research. We would like to continue our efforts to improve the coverage and quality of the Old Javanese Wordnet. This would include: adding more synsets and lemmas, adding new concepts that do not appear in the PWN and proposing them as new English entries, validating and revising the list of candidate senses generated through the methods explained in Section 4, investigating the scientific names, as well as adding example sentences for each sense. Once the Old Javanese Wordnet reaches a sufficient coverage, we would like to use it to build and develop a wordnet for Modern Javanese and to research a variety of topics, including: semantic changes in Sanskrit origin-Old Javanese words, semantic changes from Old Javanese to Modern Javanese, and verb subcategorization in Old Javanese to be used in building a computational grammar for Old Javanese (Moeljadi, 2019).

9. Acknowledgements

We gratefully acknowledge contributions made by Doug Cooper from the Center for Research in Computational Linguistics (CRCL), Bangkok and Arlo Griffiths from the École Française d'Extrême-Orient (EFEO) for providing the digitized Old Javanese–English Dictionary (OJED) data which is released under CC 4.0 (Doug Cooper, personal communication). In building the online OJED, CRCL gratefully acknowledges the cooperation and assistance of Stuart Robson who gave the project his blessing, Thomas Malten and the staff of the Sanskrit and Tamil Publishing Service (SPS) who keyed the text, Kees Waterman and the Koninklijk Instituut voor Taal-, Land- en Volkenkunde (KITLV) for granting permission to prepare the online edition, and Arlo Griffiths and the EFEO for financing rekeying of the source text. We thank Luis Morgado da Costa and Francis Bond from NTU, Singapore for answering our questions regarding wordnets. We also thank our anonymous reviewers and Arlo Griffiths for their comments on this paper. The first author gratefully acknowledges the support of the European Regional Development Fund-Project "Sinophone Borderlands – Interaction at the Edges" CZ.02.1.01/0.0/0.0/16_019/0000791.

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

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

[11] http://globalwordnet.org/

10. Bibliographical References

Acri, A. and Griffiths, A. (2014). The Romanisation of Indic Script Used in Ancient Indonesia. *Bijdragen tot de taal-, land-en volkenkunde/Journal of the Humanities and Social Sciences of Southeast Asia*, 170(2-3):365–378.

Acri, A. (2017). *Dharma Pātañjala: a Śaiva scripture from ancient Java studied in the light or related Old Javanese and Sanskrit texts*. International Academy of Indian Culture and Aditya Prakashan, New Delhi.

Aminullah, Z. P. (2019). Candrakiraṇa: Présentation du texte avec étude de sa première partie consacrée entre autres aux structures métriques. Master's thesis, EPHE, France.

Bond, F. and Foster, R. (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.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: The Collaborative Interlingual Index. In Christiane Fellbaum, et al., editors, *Proceedings of the Global WordNet Conference*, pages 50–57, Bucharest.

Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the third international WordNet conference*, pages 29–36.

Creese, H. (2001). Old Javanese Studies: A Review of the Field. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 157(1):3–33.

Dániel, B. and Griffiths, A. (2019). Transliteration Guide for Members of the DHARMA Project. (halshs-02272407v2).

David M. Eberhard, et al., editors. (2019). *Ethnologue: Languages of the World*. SIL International.

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

Gonda, J. (1952). *Sanskrit in Indonesia*. International Academy of Indian Culture, Nagpur.

Gunawan, A. (2016). Bhīma Svarga: étude d'un texte vieux-javanais et de sa transmission manuscrite, mémoire de M2, Institut national des langues et civilisations orientales. Master's thesis, INALCO, France.

Hunter, T. M. (2009). Some Problems in the Study of Old Javanese as a Linguistic System. In *The Second International Symposium On The Languages Of Java*, Lombok, Indonesia.

Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., and Bhattacharyya, P. (2010). Introducing Sanskrit Wordnet. In *Proceedings of the 5th Global Wordnet Conference (GWC 2010)*, pages 287–294, Mumbai. Narosa.

Moeljadi, D., Kamajaya, I., and Amalia, D. (2017). Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications. In Hai Xu, editor, *Proceedings of the 11th International Conference of the Asian Association for Lexicography*, pages 64–80. the Asian Association for Lexicography, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

Moeljadi, D. (2019). Building a computational grammar for Old Javanese. In *The Seventh International Symposium on the Languages of Java*, Banyuwangi, Indonesia.

Pollock, S. (1996). The Sanskrit cosmopolis, 300-1300 CE; Transculturation, vernacularization, and the question of ideology. In Jan E.M. Houben, editor, *Ideology and Status of Sanskrit: Contributions to the History of the Sanskrit Language*, chapter 10, pages 197–248. E.J. Brill, Leiden.

Robson, S. O. (2008). *Arjunawiwāha: the marriage of Arjuna of Mpu Kaṇwa*. KITLV Press, Leiden.

Sio, J. U.-S. and Morgado da Costa, L. (2019). Building the Cantonese Wordnet. In *Proceedings of the Tenth Global Wordnet Conference*, pages 206–215, Wroclaw, Poland.

Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht.

Wojowasito, S. and Mills, R. F. (1980). *A Kawi Lexicon*. University of Michigan Centers for South and Southeast Asian Studies.

Worsley, P., Supomo, S., Hunter, T. M., and Fletcher, M. (2013). *Mpu Monaguṇa's Sumanasāntaka: an Old Javanese epic poem, its Indian source and Balinese illustrations*. Brill, Leiden.

Zoetmulder, P. J. (1974). *Kalangwan: A Survey of Old Javanese Literature*. Martinus Nijhoff, 'S-Gravenhage.

Zoetmulder, P. J. (1982). *Old Javanese–English Dictionary*. Martinus Nijhoff, 'S-Gravenhage.