Experiment on building Sundanese lexical database based on WordNet

Sari Dewi Budiwati, Novihana Nurani Setiawan
Information System Study Program,
School of Applied Science, Telkom University
Jl. Telekomunikasi, Ters. Buah Batu – Bandung
saridewi@tass.telkomuniversity.ac.id

Abstract. Sundanese language is the second biggest local language used in Indonesia. Currently, Sundanese language is rarely used since we have the Indonesian language in everyday conversation and as the national language. We built a Sundanese lexical database based on WordNet and Indonesian WordNet as an alternative way to preserve the language as one of local culture. WordNet was chosen because of Sundanese language has three levels of word delivery, called language code of conduct. Web user participant involved in this research for specifying Sundanese semantic relations, and an expert linguistic for validating the relations. The merge methodology was implemented in this experiment. Some words are equivalent with WordNet while another does not have its equivalence since some words are not exist in another culture.

1. Introduction
Indonesia has 719 individual languages. Of these, 706 are living and 13 are extinct [1]. Sundanese language is the second biggest language used in Indonesia. It is used by 34,000,000 population, most of the language is used in West Java and Banten province and consist of different dialect: Sumedang, Cianjur, Ciamis, Serang, Bogor, Subang, Tasikmalaya, Purwakarta and Cirebon dialect [2]. Sundanese language is considered as stable but threatened language [2] as because of some reason:

- Sundanese language rarely used by youth
  Regional languages are considered as an old-fashioned language by Indonesian youth, this is also happening with Sundanese. They are prouder using Indonesian or English language instead Sundanese as it is one of the effects of globalization. Now, there is only 40% Sundanese youth who know and understand Sundanese language in West Java Province [3].

- Competition between regional language, national language, and foreign language
  After the Indonesian Youth pledge on October 28th, 1928, a new generation used the Indonesian language in their daily activities: school, books, news, office work, etc. Additionally, English as global language affecting the youth generation also. Due to these, many Indonesian youths tend to use Indonesian or English language than regional language causing the decreasing in regional language.

In order not to be categorized as an endangered language, there is need some efforts to preserve the language [2]:

- Using Sundanese language in the family or social environment.
This effort had been proclaimed by Bandung Mayor in 2013. Every Wednesday, Bandung citizens are encouraged to use Sundanese for communication, it is known as Sundanese Wednesday (Rebo nyunda). But this culture is done in Bandung city only, not in another area of West Java and Banten province, therefore the chance for the decreasing still can be found.

- The inclusion of Sundanese language in school curriculum
  Indonesian school curriculum was encouraged to have the local subject as a distinction with other region and to preserve regional wisdom. This also applied in West Java and Banten province. Elementary to high school had been introduced Sundanese languages in their curriculum.

- The planning of corpus can be pursued by accelerating the alignment of Sundanese language through the absorption of Indonesian vocabulary, or other languages to express concepts, especially science and another modern life.

In this paper, we described our effort to maintain and preserve Sundanese language using information technology approach. We built a Sundanese lexical database based on Princeton WordNet (2) and Indonesian WordNet (3), called Sunda WordNet. It is developed by merge approach as Sundanese synset is specified first the define its semantic relations and then mapped into WordNet synsets. The selection method was chosen based on Sundanese language characteristics that it has rich of same meaning for one word, called Code of Conduct (undak-usuk basa).

2. Sundanese language code of conduct

Sundanese language has four [5] speech level or language code of conduct [4], but almost all of Sundanese literature makes into three level:

- Ribaldry level. It is used for friends whose age is below, or same age or for a best friend.
- Standard level. It is used for his/herself
- Polite level. It is used for upper level, someone whose doesn’t know, or for someone who is being respectful.

In Table 1 given examples how different words can be used based on the language level although it has the same meaning. Some words changes are simple as we merely need to replace the words, for instance in no 2, we can replace the word "sick" for each level with "damang", "udur", and "gering", without having to change the previous sentence: "What's the matter, ..." ("Ku naon Jang,..."). But for other, it will affect the whole sentence. For example, in no 4, the word "eat" for each level are "tuang", "neda" and "dahar", but for ribaldry levels, we have to change "want to" with "arek. Therefore, Sundanese words are rich for synonym words. Some example of language code of conduct with WordNet equivalence is shown in Table 2.

| No | English sentence                        | Sundanese sentence |
|----|-----------------------------------------|--------------------|
|    |                                         | Polite             | Standard          | Ribaldry              |
| 1  | You don't hear it, do you?             | Teu kadangku keceupil anjeun? | Teu kadenge ku ceuli maneh? | Nya teu kadenge ku jegebir sia teh? |
| 2  | What's the matter, are you sick?       | Ku naon Jang, teu damang? | Ku naon Jang, udar? | Ku naon Jang, gering? |
| 3  | I am sorry, I could not come yesterday | Punten we kamari teh abdi mah teu tiasa sumping | Punten we kamari teh abdi mah teu tiasa dongkap | Punten we kamari teh abdi mah teu datang. |
Do you want to eat, mom? | Bade tuang, bu? | Bade neda, bu? | Arek dahar, bu?
---|---|---|---
1. Enter, come in, get into, get in, go into, go in, move into | Lebet | Lebet | Abus, Asup
2. Little brother | Rai, rayi | Adi | Adi
3. Compete, contend | Aben | Adu | Adu
4. Learn, larn, acquire | Wulang | Wuruk | Ajar
5. Grandfather, gramps, granddad, grandad, graddaddy, grandpa | Tuang eyang | Pun aki | Aki
6. Admit, acknowledge | Angken, ngangken | Aku, ngaku | Aku, ngaku
7. Hoop, ring | Cingcin, lelepen | Ali | Ali
8. Sick | Teu damang | Udur | Gering
9. Arrive | Sumping, rawuh | Dongkap | Cunduk, datang
10. Eat | Tuang | Neda | Dahar
11. Ear | Cepil | Ceuli | Ceuli, Jejebir

### Table 2. Sundanese language code of conduct example

| No | English words | Polite | Standard | Ribaldry |
|----|---------------|--------|----------|----------|
| 1  | Enter, come in, get into, get in, go into, go in, move into | Lebet | Lebet | Abus, Asup |
| 2  | Little brother | Rai, rayi | Adi | Adi |
| 3  | Compete, contend | Aben | Adu | Adu |
| 4  | Learn, larn, acquire | Wulang | Wuruk | Ajar |
| 5  | Grandfather, gramps, granddad, grandad, graddaddy, grandpa | Tuang eyang | Pun aki | Aki |
| 6  | Admit, acknowledge | Angken, ngangken | Aku, ngaku | Aku, ngaku |
| 7  | Hoop, ring | Cingcin, lelepen | Ali | Ali |
| 8  | Sick | Teu damang | Udur | Gering |
| 9  | Arrive | Sumping, rawuh | Dongkap | Cunduk, datang |
| 10 | Eat | Tuang | Neda | Dahar |
| 11 | Ear | Cepil | Ceuli | Ceuli, Jejebir |

### 3. Princeton WordNet

We take WordNet definition from [6] because our goal in this research is to collect Sundanese words and classified as lexical and semantic relations. They stated that "Princeton WordNet, or simply PWN, is a lexical resource for the English language containing a large database of open-class words, i.e. nouns, verbs, adjectives, and adverbs. These words are clustered together based on their meaning into synonym sets, or synsets. Each instance of a word in a synset is marked as a separate sense of that word. PWN further specifies semantic relations between synsets and specific word senses, among others antonym, hypernym, and meronym" [6].

Further definition of relations in WordNet are described below:

- **Lexical relations, consist of synonym and antonym**
  - **Synonym:**
    Two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made [7]. For synonym in Sundanese language is explained in table 1.
  - **Antonym:**
    The antonym of a word x is sometimes not-x, but not always [7]. Like in English, "not" (henteu/teu) can be placed before words, for Sundanese generally the same. For example, in the figure below, "sick" for Sundanese is "teu damang", meanwhile "healthy" for Sundanese is "damang". But like [7] said, it is not always "not-x", for example “in” in Sundanese is "lebet", meanwhile "out" for Sundanese is "kaluar".
b. Semantic relations, consist of Hyponym/Hypernym and Meronym/Holonym
   - Hyponymy/hypernym is a subordination/superordination, subset/superset or IS A relation [7]. Sundanese example for hyponymy/hypernym is shown in figure below

   ![Figure 2. Hyponym/Hypernym example](image)

   - Meronym/Holonym is a part-whole (or HAS A) relation [7]. Sundanese example for meronym/holonym is shown in figure below

   ![Figure 3. Meronym/Holonym example](image)

4. Related Work
The Sundanese language research in Natural Language Processing field still minor, this is shown as there's still few research publications and basic resources tools in Sundanese for advanced end-user technologies, such as corpora, part of speech, parser, etc. Therefore, Sundanese is considered as under resource language. The first research is The Crubadan Project, where one of its research resources is Sundanese lexicons. The Crubadan project is developed by a web crawler, using simple character trigrams for language recognition. They collect the lexicon from 330 web pages and resulted in 50,000 words, 50,000 words bigram and 9,520 chartigrams [8].

5. Methodology and approach
There are two general approaches to construct a WordNet [6]:
   - The expand model
     The synsets in WordNet are translated to the target language, afterward, its semantic relationships holding between synsets, are also transferred over.
   - The merged model
     The synsets and semantic relationship are independently specified, afterward, its mapped into WordNet synsets through the definition of an equivalence relation.

Meanwhile, another WordNet research language use a hybrid approach [9].
a. Design and Implementation

Sunda WordNet is developed by merge approach. We choose this method as Sundanese language is rich with synonyms. In this research, first Sundanese synset is specified, further define its semantic relations, and then mapped into WordNet synset (Figure 4). As for implementation design, we choose use-case diagram to describe whole system Figure 5.

![Figure 4. Merge approach for experiment](image1)

**Figure 4. Merge approach for experiment**

**Figure 5. Use-case diagram process**

i. **Specified Sundanese synset.** Sundanese synset is easier to be defined since it is based on Sundanese language code of conduct. In Figure 6, we use a random number for synonym ID and we collect synonym words (synset) from [10] which consist of 2.259 words. Meanwhile, we collect the description for definition column from [11]. Last, example column and idKelasKata is added by participant users. idKelasKata is a field to define whether the synset is noun, verb, and adverb.

![Figure 6. Example of synonym table](image2)

**Figure 6. Example of synonym table**

ii. **Define Sundanese semantic.** Lexical antonym was defined from the same source [10] which results as 94 pair synonym-antonym. Since the result is still a very small number, we added from [12] as we show some example in Table 3. We change the synset_ID from a random number in Figure 6 to be less simple as seen in Table 3.

| Id | Synset_ID | antonyms_ID | definition | example | idKelasKata |
|----|-----------|-------------|------------|---------|-------------|
| 1  | 4G29B18A49d9113C0146d91720a0001 | Kupala | Srahna Iedagor kana ilang | K001 |
| 2  | 4G29B18A49d9113C0146d91720a0001 | Gara | Duda Afoya bu sari goa nga masikina koi auki |
| 3  | 4G29B18A49d9113C0146d91720a0001 | Njul | Maneh makt geda hulu K001 |
| 4  | 4G29B18A49d9113C0146d91720a0001 | Duariat | Kupala (bahasa kunan) | K001 |
| 5  | 4G29B18A49d9113C0146d91720a0001 | Pans | Makte, kepilai ati pengmahat | K001 |
| 6  | 4G29B18A49d9113C0146d91720a0001 | Sosa | Kars, kepilai ati pengmahat | K001 |
| 7  | 4G29B18A49d9113C0146d91720a0001 | Kuris | Rangwil (ang bumbah) | K001 |

**Table 3. Sundanese antonym mapping**
For semantic relations, we involved web user participant due to no resource or dictionary of Sundanese semantic words. Another objective is to promote and actively participate from Sundanese people. In Figure 7, we design the graphical user interface for the participant to choose the relation between synset, whether it is holonym-meronym or hypernym-hyponym. The example result of the relations is shown in Table 4 and Table 5.

### Table 4. Sundanese hypernym-hyponym mapping

| HH_ID | Parent_ID | Words  | Sinonym_ID | Words      |
|-------|-----------|--------|------------|------------|
| HH_01 | S09       | Kembang| S10        | Kembang-Eros |
| HH_02 | S09       | Kembang| S11        | Malati     |

### Table 5. Sundanese holonym-meronym mapping

| HM_ID | Parent_ID | Words | Sinonym_ID | Words |
|-------|-----------|-------|------------|-------|
| HM_01 | S12       | Sira-h| S18        | Panon |
| HM_02 | S12       | Sira-h| S21        | Bengeut |

Choosing the synset relation will be difficult for some people because not all Sundanese people understand the synset words. As the result, many of participant gives wrong answer and it will be affected the result. Therefore we provide a validator user who’s a Sundanese expert language which will approve the answer (Figure 8).

### Figure 7. Synonym mapping to meronym and holonym

### Figure 8. Validation page

**iii. Sundanese synset mapping into WordNet synset.** Next step is mapping the Sundanese synset into WordNet synset. The mapping process is done by sense of understanding from expert to minimize the error. We use WordNet description from Indonesian WordNet [6] and [7] as shown in Figure 9 and then define the Sundanese synset as shown in Figure 10. Some examples of mapping result are shown in Table 6 and Table 7. Meanwhile, the entity relationship diagram for the whole system is shown in Figure 11.
In this step, we found words which still does not have its equivalence in WordNet. For example:

- **Ngaso (aso, leleson)** is a Sundanese word for having rest-activity from works. We found its equivalence as a “break” words, but after examining from WordNet synset, we concluded it is still not best equivalence or best answer.

- **Karembong (kekemben)** is a Sundanese word for clothes which is wrapped around the body, yet we do not find its equivalence in WordNet.

The example above is some words as a result of a culture, as explained by [4]:

"On the other hand, what is seen in a language will be reflected in its culture. For example, the British and other European people, who are not familiar with eating rice, only have one word, rice, to express the concept of gabah (rice in the husk), beras (uncooked rice with the husk removed), and nasi (cooked rice)." (Sudaryat, 2015)
6. Evaluation and future research
In this experiments the result data still rough due to several reasons
   1. Participant involvement for specifying semantic relation is helpful but causing some problems, for example, many participants gives false answer, therefore validator have to check carefully. Another problem rise because it takes a lot of time due to its voluntary participation. In future research, this would might be done with simple Cartesian Product algorithm.
   2. Since synset mapping from Sundanese synset to WordNet is done by expert manually and it needs careful attention, for future research this step might be done using the Sundanese-English dictionary as a seed to mapping the synset. In addition, web crawling can be used to collect the words or parallel sentences from web page and mapping it to WordNet using dictionary.
   3. Although we already define three words category (noun, verb and adverb) but it still an issue in this experiment since it is use participant decision which caused same problem in (1). Sundanese-English dictionary can be use in this step since most of dictionary already define whether the words is noun, verb or adverb.

7. Conclusion
Merge approach is suitable in this experiments since Sundanese has language code of conduct, which consists of three speech levels of words for same meaning. We collect 2.259 words from language code of conduct and is used as base for synset. Next, specifying semantic relations need more effort because it is still based on user's participation. In this experiment, we found words which are still not available its equivalence in WordNet, due to language is a result of a specific culture.
References

[1] Simons GF, Fennig CD. Ethnologue: Languages of the World, Twentieth edition Dallas, Texas: SIL International; 2017.

[2] Sobarna C. Bahasa Sunda Sudah Di Ambang Kematiankah? HUBS Asia. 2007 June; 11: p. 13-17.

[3] Mulyanah A. Republika Online. [Online].; 2013 [cited 2017. Available from: http://nasional.republika.co.id/berita/nasional/jawa-barat-nasional/13/08/26/ms4nkwbahasa-sunda-terancam-punah.

[4] Sudaryat Y. The Mindset of Sundanese People in Their Language Expressions. Indonesian Journal of Applied Linguistics. 2015 July; 5: p. 126-133.

[5] S A. Tatakrama (Undak-Usuk) Basa Sunda: Bandung: Pustaka Adhigama; 1997.

[6] Putra DD, Arfan A, Manurung R. Building an Indonesian WordNet.. In In Proceedings of the 2nd International MALINDO Workshop; 2008.

[7] Miller A, Beckwith , Fellbaum , Gross , Miller. Introduction to WordNet: An On-Line Lexical Database. International Journal of Lexicography. 1993.

[8] Scannell P. The Crubadan Project: Corpus building for under-resourced languages. In In Proceedings of the 3rd Web as Corpus Workshop; 2007; Louvain-la-Neuve, Belgium.

[9] Prabhu , Desai , Redkar , Prabhugaonkar , Nagvenkar , Karmali R. An Efficient Database Design for IndoWordNet Development. In In COLING 2012 for 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP); 2012.

[10] Haerudin D. Indonesia University of Education. [Online].; 2012 [cited 2014. Available from: http://file.upi.edu/Direktori/FPBS/JUR._PEND._BAHASA_DAERAH/196408221989031-DINGDING_HAERUDIN/

[11] Danadibrata RA. Kamus Basa Sunda Bandung: PT. Kiblat Buku Utama; 2015.

[12] Tamsyah BR, Purmasih P, Supratman. Kamus Sinonim-Antonim Bahasa Sunda Bandung: Pustaka Setia; 1998.