Building the KamusBesarBahasa Indonesia (KBBI) Database and Its Applications

David Moeljadi
Nanyang Technological University
D001@e.ntu.edu.sg

Ian Kamajaya
ASTrioPte Ltd
ian@astriotech.com

Dora Amalia
Badan Bahasa
dora.amalia@kemdikbud.go.id

Abstract

The official dictionary of the Indonesian language, KamusBesarBahasa Indonesia (KBBI), is published by BadanPengembangandanPembinaanBahasa (The Language Development and Cultivation Agency) or BadanBahasa, under the Ministry of Education and Culture of the Republic of Indonesia. The fourth edition of KBBI (Sugono 2008) has more than 92,000 entries and 100,000 senses and contains a wealth of linguistic information and cultural diversity of Indonesia. However, the data was available only in Microsoft Excel and Word files in exactly the same format as the one in the printed dictionary. Its online edition was only meant for basic word search by entry words. Thus, in order to create an online dictionary application which has advanced search capabilities, building a database is very vital: the data structure needs to be identified and the data itself needs to be cleaned so that it can be broken down based on its components. Atkins and Rundell (2008: 114) state that a database is one of the three main components of Dictionary Writing System (DWS). This paper describes our efforts in building the KBBI database in SQLite (www.sqlite.org) using Python programming language (www.python.org) and presents some applications for lexicographic and linguistic research and analysis. The KBBI database is employed for the online DWS application called KBBI DalamJaringan or KBBI Daring (https://kbbi.kemdikbud.go.id) (Kamajaya et al. 2017), the offline KBBI mobile applications in Android and iOS, and the printing of the latest, fifth edition of KBBI (Amalia 2016).

Keywords: *KBBI, database, Indonesian language dictionary, machine-tractable dictionary*

1. Introduction

KamusBesarBahasa Indonesia (KBBI) is the official dictionary for Indonesian,[1]

[1] Indonesian (ISO 639-3: ind), called *bahasa Indonesia* (lit. "the language of Indonesia") by its speakers, is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties (Lewis 2009). It is spoken mainly in the Republic of Indonesia as the

published by BadanPengembangandanPembinaanBahasa (The Language Development and Cultivation Agency) or BadanBahasa under the Ministry of Education and Culture of the Republic of Indonesia. Up until present, KBBI is the most comprehensive and the most authoritative reference for the Indonesian language. The first edition of KBBI, published in 1988, has 62,000 entries. The number of entries increased to 72,000 or about 10,000 entries over three years in the second edition (1991). The third edition of KBBI, published in 2001, contains 78,000 entries and seven years later, the fourth edition of KBBI's number of entries increased to more than 92,000. The latest, fifth edition of KBBI was released for the first time in 2016 in three formats: printed, online, and offline versions. These three versions are launched to meet the needs of all users. Figure 1 shows the printed version of KBBI from the first edition to the fifth edition. This paper describes our work in 2016 on making a database for the fourth edition of KBBI which is then employed for the printed, online, and offline versions of the fifth edition of KBBI.

Regarding the online KBBI before 28 October 2016, it used the data from the third edition of KBBI and allowed searches only by headwords. The search results were presented exactly in the same format as the one in the printed version, i.e. using bold or italic typefaces and different punctuations, such as colons and semicolons (see Section 2 for the details of the formatting effects). These formatting effects serve only as stylistic presentations and do not distinguish the fields or their structure explicitly. For example, to look up mengacang, a user must first look up the root word (kata dasar) kacang, as shown in Figure 2. This may present some difficulties if the user is not familiar with Indonesian morphological rules. The users cannot perform more targeted searches and computer applications cannot utilize the data fully. This can be overcome by identifying the data structure, cleaning the data, and breaking it down based on its components or structure of dictionary entries.

We identify the data structure and break it down using regular expressions in Python programming language (www.python.org). The results are converted into a SQLite database (www.sqlite.org) to facilitate more specific and targeted word lookup and analysis. Lim et al. (2016) mention the categorization of lexical resources in terms of their digital readiness for natural language processing (NLP) work, from paper dictionaries, machine-readable dictionaries, machine-tractable dictionaries, to semantic rich resources. Paper dictionaries are traditional dictionaries printed on paper. They are only for human consumption. The contents are presented with text formatting effects and organized by headwords. Machine-readable dictionaries (MRDs) are digitized versions of the original paper-printed versions and are the most common form of electronic dictionaries, which retain the text formatting styles. The previous online KBBI, as shown in Figure 2, was an MRD. Machine-tractable

---

sole official and national language and as the common language for hundreds of ethnic groups living there (Alwi et al. 2014: 1-2). In Indonesia it is spoken by around 43 million people as their first language and by more than 156 million people as their second language (2010 census data). The lexical similarity is over 80% with Standard Malay (Lewis 2009). It is written in Latin script.

dictionaries are MRDs with machine-tractable structures, i.e. all fields and hierarchy of the entries are specifically marked and delineated, such that different information can be identified and extracted. Our work was to bring the KBBI to the level of this digital-readiness. Semantic rich resources are machine-tractable dictionaries with semantic information for each sense entry. They are very useful for NLP tasks, such as text categorization, sentiment analysis, and information extraction. However, this is outside the scope of our work.

Figure 1 KamusBesarBahasa Indonesia (KBBI), from the first to the fifth edition

Figure 2 Screenshot of the online KBBI before 28 October 2016

2. The KBBI dictionary format

KBBI is a general dictionary whose macrostructure has a hierarchical order. The schematic is arranged by placing the basic form (the root word or kata dasar) as the headword or the lemma. The information fields in an entry structure include a headword or lemma; variant forms; pronunciations; labels: parts-of-speech, styles, languages, domains, idioms, abbreviations; sublemmas/subentries: derived words, multiword expressions (MWEs) including compounds, idioms, and proverbs; definitions; cross-references; examples; scientific names; and chemical formulas.

Figure 3 shows us that the headword or the lemma is in bold type with periods for syllabification, followed by the pronunciation, surrounded by slashes. The part-of-speech label is written in italic type, following the

pronunciation. If there is more than one definition phrase in one sense, the definition phrases are separated by semicolons. If there is an example, a colon is put after the last definition and followed by a space; the example is in italic type. The lemma is represented by two hyphens in the example. If a lemma is an abbreviation, a label for abbreviation is placed before the definitions. If there is more than one sense, a polysemy number is written in bold type before each sense and senses are separated by semicolons. If the definition is in a foreign language, it is written in italics. Figure 4 illustrates these formatting effects. If a lemma has a chemical formula or a scientific name, it is written after the definition and preceded by a semicolon, as shown in Figure 5. The scientific name is written in italic type. Some numbers in the chemical formula are subscripted.

If a lemma is homonymous, a homonymy number is placed before the lemma in superscript bold type. If a particular label is appropriate for every sense, it is written before the first polysemy number. If a label is appropriate only for a particular sense, it is written after the respected polysemy number for that sense. Subentries, such as compounds and derived words, are in bold type. Subentries are separated by semicolons. If a derived word has an example, it is represented by a tilde in the example. Figure 6 shows these formatting effects. There is a special feature which distinguishes KBBI from other monolingual dictionaries, i.e. the order of the derived words is not arranged alphabetically, but in accordance with the paradigm of word formation. For example, tinju“boxing”, as a lemma, has meninju “to box” as a transitive verb followed by peninju “boxer”, peninjuan “act/process of boxing”, and tinjuan“the result of boxing”. This sequence of verbs, actors, acts/processes, and results is called the paradigm of word formation.

If a lemma or a sublemma appears in proverbs or idioms, the lemma is represented by two hyphens, while the sublemma is represented by a tilde, same as in the example field. Both the proverbs and the idioms are in italics, followed by a comma and pb for proverbs (pb stands for peribahasa) or ki for idioms (ki stands for kiasan), as shown in Figure 7. For cross-references, if a lemma is non-standard, a right arrow is placed after it, followed by the standard lemma in bold type. If a lemma is a part of an idiomatic compound, it is followed by lihat “see” and the cross-referenced lemma printed in bold type (see Figure 8). Up until the fourth edition of KBBI, the dictionary data with the formatting effects mentioned above was available only in Word and Excel files. The following section describes our work in breaking down the components based on the formatting effects.

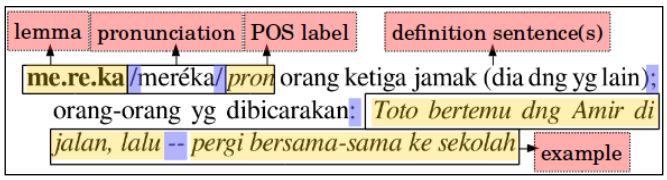


Figure 3 Example entry mereka “they”

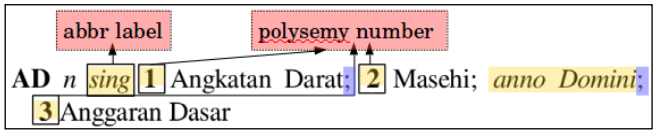


Figure 4 Example entry AD

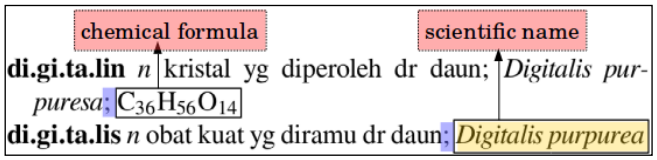


Figure 5 Example entries digitalin and digitalis

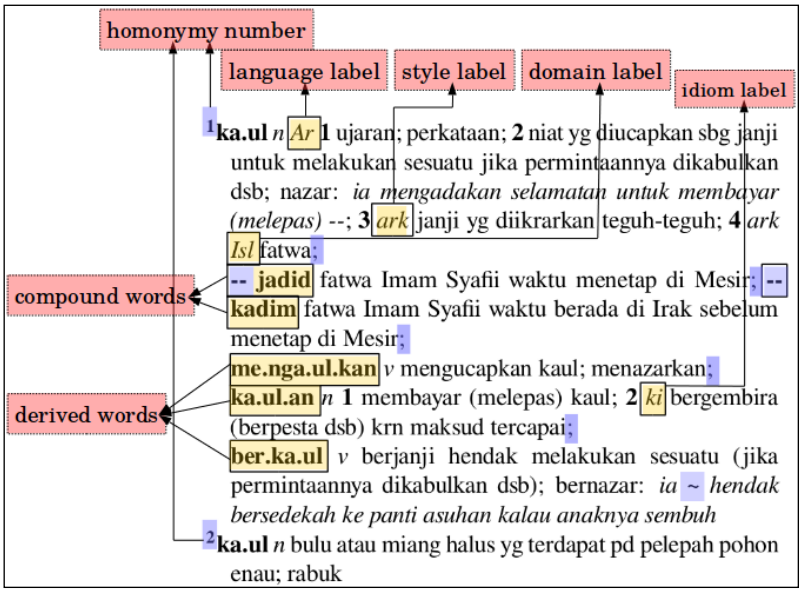


Figure 6 Example entrieskaul “vow”

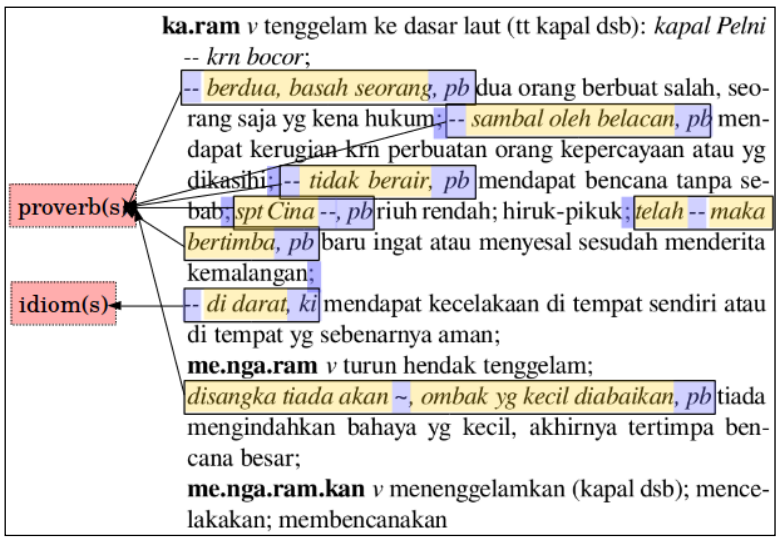


Figure 7 Example entry karam “shipwrecked”

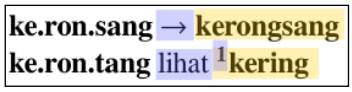
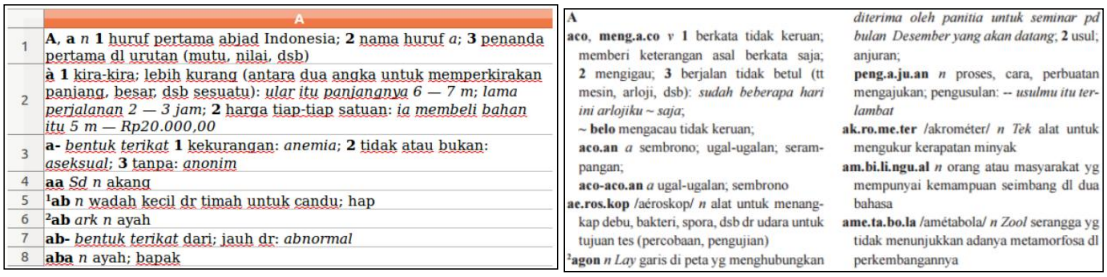


Figure 8 Example entrieskeronsang and kerontang

3. Cleaning-Up, Conversion, and Database Creation

The cleaning-up and conversion processes of data in Word and Excel files are quite tricky. This is because the data available in Word and Excel are formatted text (i.e. some of the data are in bold type, some others are in italic type, some others are superscripted, subscripted, or any combination of those, as described in Section 2) and that we want to keep the format as we transfer the data into the database. Figure 9 shows a part of the Word and Excel files. The format, not just the text, is a part of the lexicographic information and must not be removed during the conversion process. Hence, simple data extraction from Excel cells and Word paragraphs to the database entries cannot be done as it would not retain the text format.

In order to retain the text format in the conversion, the formatting effects in the Word and Excel files must be read, too. Therefore, a Windows Form application named KBBICleaner(see Figure 10) is created using .Net Framework to help us complete the task. The program uses Word-and-Excel-compatible Microsoft-created dynamic link libraries (.dll), namely Microsoft.Office.Interop.Word and Microsoft.Office.Interop.Excel, to extract the data from Word and Excel in the Rich Text Format (RTF). Figure 11 shows a part of the RTF file. Furthermore, to ease the cleaning-up process, the program is designed with three additional main functionalities: (a) File and string manipulation, (b) List of text filter and conversion, and (c) List of regex filter and conversion, explained inthe following subsections.



	A
1	**A, a** *n* **1** huruf pertama abjad Indonesia; **2** nama huruf *a*; **3** penanda pertama dl urutan (mutu, nilai, dsb)
2	**à 1** kira-kira; lebih kurang (antara dua angka untuk memperkirakan panjang, besar, dsb sesuatu): *ular itu panjangnya 6 — 7 m; lama perjalanan 2 — 3 jam*; **2** harga tiap-tiap satuan: *ia membeli bahan itu 5 m — Rp20.000,00*
3	**a-** *bentuk terikat* **1** kekurangan: *anemia*; **2** tidak atau bukan: *aseksual*; **3** tanpa: *anonim*
4	**aa** *Sd n* akang
5	**[1]ab** *n* wadah kecil dr timah untuk candu; hap
6	**[2]ab** *ark n* ayah
7	**ab-** *bentuk terikat* dari; jauh dr: *abnormal*
8	**aba** *n* ayah; bapak

aco, meng.a.co *v* **1** berkata tidak keruan; memberi keterangan asal berkata saja; **2** mengigau; **3** berjalan tidak betul (tt mesin, arloji, dsb): *sudah beberapa hari ini arlojiku ~ saja*;
~ **belo** mengacau tidak keruan;
aco.an *a* sembrono; ugal-ugalan; seram-pangan;
aco-aco.an *a* ugal-ugalan; sembrono
ae.ros.kop /aéroskop/ *n* alat untuk menang-kap debu, bakteri, spora, dsb dr udara untuk tujuan tes (percobaan, pengujian)
[2]agon *n Lay* garis di peta yg menghubungkan

diterima oleh panitia untuk seminar pd bulan Desember yang akan datang; **2** usul; anjuran;
peng.a.ju.an *n* proses, cara, perbuatan mengajukan; pengusulan: *-- usulmu itu ter-lambat*
ak.ro.me.ter /akrométer/ *n Tek* alat untuk mengukur kerapatan minyak
am.bi.li.ngu.al *n* orang atau masyarakat yg mempunyai kemampuan seimbang dl dua bahasa
ame.ta.bo.la /amétabola/ *n Zool* serangga yg tidak menunjukkan adanya metamorfosa dl perkembangannya

Figure 9 A part of KBBI Fourth Edition in Microsoft Excel and Word files

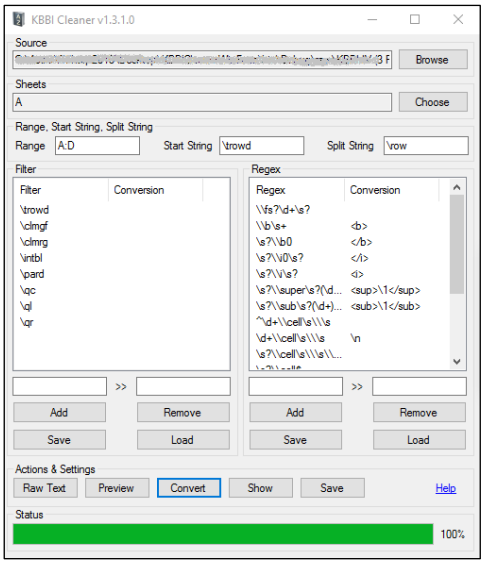


Figure 10 Screenshot of KBBICleaner

```
\trowd \trgaph30\trleft-30\trrh317\cellx1040\clmgf \cellx2351
\qc \f5\fs22 \cf55 1\cell \ql \f6\fs22 \b A\f5\fs22 \b0 , \f6
\fs22 \i0  \f6\fs22 \b 1\f5\fs22 \b0  huruf pertama abjad Ind
\b0  nama huruf \f7\fs22 \i a\f5\fs22 \i0 ; \f6\fs22 \b 3\f5\
urutan (mutu, nilai, dsb) \cell \qr \f0\fs22 \cell
\pard \intbl \row\trowd \trgaph30\trleft-30\trrh317\cellx1040
\cellx18546\pard \intbl \qc \f5\fs22 2\cell \ql \f6\fs22 \b \
\b 1\f5\fs22 \b0  kira-kira; lebih kurang (antara dua angka u
besar, dsb sesuatu): \f7\fs22 \i ular itu panjangnya 6 \u8212
\i lama perjalanan 2 \u8212\'97 3 jam\f5\fs22 \i0 ; \f6\fs22
tiap satuan: \f7\fs22 \i ia membeli bahan itu 5 m \u8212\'97
\qr \f0\fs22 \cell
```

Figure 11 A part of the RTF file of KBBI

3.1 File and String Manipulation

This function helps us determine which portions of the Excel file to be cleaned using the program. This is used primarily for cleaning up the Excel file as it contains multiple sheets with different (inconsistent) “Range” to be cleaned up (for example, in one sheet, there might be three columns of data while in another sheet there might be four columns). It also helps us determine which starting and ending strings (in RTF format) can be used as division of cells, lines, or paragraphs.

3.2 List of Text Filter and Conversion

This function helps us process the file with a list of text filter and conversion. Some data in the RTF format are not needed (for example, the header of the file and the unused format code) and some need to be changed (for example, indicators of bold type, italic type, superscript, and subscript formats). Thus, a list of text filter and conversion will greatly help us process such data. All entries in the list will be applied to the original text in a sequential fashion, i.e. from the top entry to the bottom entry.
If a user needs to remove a certain consistently unused string, he or she needs to specify it in the filter with no conversion value. In addition, if a user needs

to replace a certain consistently appearing string, he or she needs to specify it and the desired conversion value in the list. Moreover, if a user needs to change two or more different formats into a single final format, he or she can exploit the sequential behavior of the filter to convert the earlier format(s) to the uniformed format in a sequential fashion and convert the uniformed format to the desired (single) final format.
This list of filter can be applied with both (real-time) user input values and predefined, loaded .txt file containing the filter information to further help a user save his or her filter midway whenever he or she finds the list of filter non-final and wants to continue to do it conveniently next time.

3.3 List of Regular Expression Filter and Conversion

Similar to the list of text filter and conversion above, this function helps a user with a list of regular expression (regex) -instead of text- filter and conversion. Regular expression or regex is a language for specifying text search strings which requires a pattern that we want to search for and a corpus of texts to search through (Jurafsky and Martin 2009). The regex filter behaves the same way as the text filter: it obtains and converts the filtered text according to the given list in a sequential fashion. However, it filters and converts the filtered text using regex patterns instead of doing direct conversion. Thus, this function can simply be perceived as a more powerful version of its text filter counterpart.

Naturally, however, being made of a set of regexes, this regex filter and conversion is significantly slower than the text filter and conversion. For cleaning-up process of a text data as large-sized as dictionary data, the time difference can be significant. Thus, this function is meant to help us process unused or to-be-converted data which form certain patterns. For statically written data, although they can be processed by this filter, they should be efficiently processed using the text filter instead of this regex filter.Table 1 shows some examples of the conversion from Excel to RTF and Hypertext Markup Language (HTML) and Figure 12 shows a part of the HTML file, as a result of the filter and conversion process using KBBICleaner.

3.4 Cleaning-up

After we converted the RTF file to a HTML file using KBBICleaner, we found some inconsistencies in the formatting effects and we did some cleaning-up for the data.We observe that these inconsistencies in formatting are sporadic and are due to the manual formatting work by hand. Table 2 shows some of the inconsistencies we found.In addition, we modified some definitions in order to make the formatting more consistent and to extract more information, such as chemical formulas, scientific names, and examples. Some examples are shown in Table 3.

3.5 Breaking down the components and creating a database

We wrote a Python script to break down the components or fields for each dictionary entry based on the patterns and formatting effectsdescribed in

Section2, using regex. Figure 13 illustrates the algorithm we used to extract a number of fields in an entry. To facilitate easier manipulation of the data, all broken-down components such as lemmas, definitions, and examples were exported to a SQLite database.

3.6 Dictionary data structure

The data structure of KBBI consists of four types of data: entry, sense, example,and category. The relationship between entry and sense, as well as the one between senseand example are one-to-many. The category is a list of descriptions or a metadata forentry, sense, and example.Figure 14 illustrates the KBBI data structure. An entry can be a fixed expression (ungkapan)or a root word (kata dasar). A fixed expression should have at least one sense andone example. In this case, one fixed expression may have one to multiple senses and onesense may have one to multiple examples. A root word should have at least one cross-reference,one sense, one compound, or one derived word. In this case, one root word mayhave zero to multiple senses and one sense may have zero to multiple examples. A rootword may also have variant(s), proverb(s), and idiom(s). A proverb or an idiom shouldhave at least one sense. A compound should have at least one cross-reference or one sense.One sense may have zero to multiple examples. Similar to the root word, a derived wordshould have at least one cross-reference, one sense, or one compound. It may also havevariant(s), proverb(s), and idiom(s). The root word can be in the form of compound if itcan be affixed and have derived word(s).

Table1 Some examples of the conversion

Field	Excel	RTF	HTML
Lemma	A, a	\b A\f5\fs22 \b0 , \f6 \fs22 \b a \f7\fs22 \b0	A, a
Label	n	\i n\f5\fs22 \i0	n
Homonymy number and lemma	^{1}ab	\b \super 1\f6\fs22 \nosupersubab\f5\fs22 \b0	ab (1)
Chemical formula	Cu_3	Cu\f9\fs22 \sub 3\f5\fs22 \nosupersub	Cu₃

Table 2 Some inconsistencies in the KBBI format and the cleaning-up process

Type	Example	
	Before cleaning-up	After cleaning-up
incomplete syllabification	(ke)ro.boh.an	(ke.)ro.boh.an
a semicolon should be a colon before an example	...pangkatdsb);~nyasb gduta...	...pangkatdsb):~nyasb gduta...
a comma should be a semicolon separating examples	...spt air mengalir: --udara, -- lalulintas;...	...spt air mengalir: --udara;--lalulintas;...

pronunciations should precede labels	... n /gad ə̀n/belanak	... /gad ə̀n/ nbelanak
ki should be written in an idiom	tidak -- tidakbukan, yg...	tidak -- tidakbukan, kiyg...
a comma should precede pb in a proverb	...taklapukolehhujanpb ...	...taklapukolehhujan,pb</ i> ...
Scientific names should not be put inside brackets	...; (Aquilariamalaccensis);	...; Aquilariamalaccensis ;

Table 3 Some modifications in the KBBI definitions

(1) Move chemical formulas to the end of the definitions, preceded by a semicolon	
Before	nit.ratnKimgaramasamnitratHNO₃ ,dipakai dl campuranpupuk
After	nit.ratnKimgaramasamnitrat, dipakai dl campuranpupuk; HNO₃
(2) Change rumuskimia “chemical formula” to a semicolon	
Before	kam.fa.nanKimkristal...pdsuhu 158–159 °C danrumuskimiaC₁₀H₁₈
After	kam.fa.nanKimkristal...pdsuhu 158–159 °C; C₁₀H₁₈
(3) Move scientific names to the end of the definitions, preceded by a semicolon	
Before	tal (1)n1tumbuhanpalem;Borassusfiabelliferda unnya...; batanglontar;
After	tal (1)n1tumbuhanpalem,daunnya...; batanglontar; Borassusfiabellifer;
(4) Change msl “e.g.” before examples to a colon	
Before	(dl bentuk kata kerjaber-...-an); mslbersikutat, berkutat-kutatan;bersipandang, berpandang-pandangan
After	(dl bentuk kata kerjaber-...-an):bersikutat, berkutat-kutatan;bersipandang, berpandang-pandangan

```
<b>A, a</b><i>n</i><b>1</b> huruf pertama abjad Indonesia; <b>2</b>
nama huruf <i>a</i>; <b>3</b> penanda pertama dl urutan (mutu,
nilai, dsb)
<b>à</b> <b>1</b> kira-kira; lebih kurang (antara dua angka untuk
memperkirakan panjang, besar, dsb sesuatu):<i>ular itu panjangnya 6
\u8212\'97 7 m</i>;<i>lama perjalanan 2 \u8212\'97 3 jam</i>; <b>2</
b> harga tiap-tiap satuan:<i>ia membeli bahan itu 5 m \u8212\'97
Rp20.000,00</i>
<b>a-</b><i>bentuk terikat</i><b>1</b> kekurangan:<i>anemia</i>; <
b>2</b> tidak atau bukan:<i>aseksual</i>; <b>3</b> tanpa:<i>anonim</
i>
<b>aa</b><i>Sd</i><i>n</i> akang
<b>ab (1)</b><i>n</i> wadah kecil dr timah untuk candu; hap
<b>ab (2)</b><i>ark</i> <i>n</i> ayah
<b>ab-</b><i>bentuk terikat</i> dari; jauh dr:<i>abnormal</i>
<b>aba</b><i>n</i> ayah; bapak
```

Figure 12 A part of the HTML file of KBBI

```
For each entry in each line in the HTML file,
    if <b> is at the beginning of that line,
        extract the lemma between <b> and </b>
        if there is an opening bracket and a closing bracket in the lemma,
        or if there is a comma in the lemma,
            extract the variant form(s)
        if there is a slash after </b>,
            extract the pronunciation(s) between slashes
            if <i> appears after the second slash,
                extract the label(s) for POS, language, domain, etc.
            if there is a number surrounded by <b>…</b>,
                split and extract the senses
                for each sense,
                    if <i> appears after </b>,
                        extract the label(s)
                    extract the definition
                    if <i> appears after a colon and </i> is at the end,
                        extract the example(s)
                    if <i> appears after a semicolon and </i> is at the end,
                        extract the scientific name(s)
                    if there is some chemical elements,
                        extract the chemical formula(s)
        if there is an arrow,
        or if there is lihat,
            extract the cross-reference
```

Figure 13 A part of the algorithm used to extract a number of fields in an entry

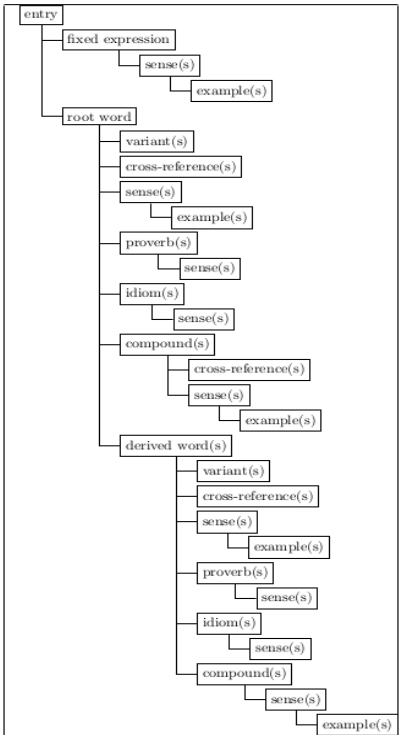


Figure 14 The KBBI data structure

4. The current state of the KBBI database and its applications

After the data was broken down into its components, we can check how many each component is in the database. As of 15 May 2017, the KBBI database contains:

- 48,142 root words (kata dasar)
- 26,198 derived words (kata turunan)
- 30,375 compounds (gabungan kata)
- 2,040 proverbs (peribahasa)
- 267 idioms (kiasan)
- 126,635 definitions (makna)
- 29,255 examples (contoh)

There are many applications can be made possible using the KBBI database. This section will provide some examples of those applications, especially for lexicography and linguistics field.

4.1 Targeted lookups

A user can search for all definitions for a word which may originate from two different headwords, e.g. mereka, using the following search procedure. The results are shown in Table 4.

```
SELECT entri, jenis, induk, lafal, kelas, makna FROM baseview WHERE entri="mereka";
```

The task of looking up phrases and MWEs such as idioms and proverbs is also made simpler, as a user would no longer need to find out which headword to look up first, e.g. the following search procedure can be used to lookup a proverb sediapayungsebelumhujan (the headword is payung ‘umbrella’). Table 5 shows the result.

```
SELECT entri, jenis, makna FROM baseview WHERE entri="sediapayungsebelumhujan";
```

Linguists and etymologists can also search specific entries by their labels. For example, a user can search archaic (ark) lemmas originating from Javanese (Jw) using the following search procedure. Table 6 shows the search results.

```
SELECT entri, ragam, bahasa, makna FROM baseview WHERE ragam="ark" and bahasa="Jw";
```

Table 4 Search results for all definitions of words with orthographic form mereka

Entri (entry)	Jenis (type)	Induk (root)	Lafal (pronunciation)	Kelas (wordclass)	Makna (definition)

merek a	root	(null)	mer ẻka	pron	orangketigajamak (diadngyg...
merek a	derive d	reka	(null)	v	menyusun (mengatur, ...
merek a	derive d	reka	(null)	v	mencariakal (ikhtiar, daya…
merek a	derive d	reka	(null)	v	memikirkan (sesuatu);...
merek a	derive d	reka	(null)	v	membayangkan (dl angan-...
merek a	derive d	reka	(null)	v	menduga; mengira-ngirakan

Table 5 Search result for the proverb sediapayungsebelumhujan

Entri (entry)	Jenis (type)	Makna (definition)
sediapayungsebelumhujan	proverb	bersiapsediasebelumterjadiygkurangbaik

Table 6 Search results for all lemmas with labels ark (archaic) and Jw (Javanese)

Entri (entry)	Ragam (style)	Bahasa (language)	Makna (definition)
cutel	ark	Jw	tamat; habis (ttceritadsb); berakhir
gundang	ark	Jw	lekum; tenggorok
pembarap	ark	Jw	anaksulung
sikep	ark	Jw	orangdrdesaygmempunyaikewajibanmelakukan ...
ubel-ubel	ark	Jw	tentaraInggrisasal India
wiyata	ark	Jw	pengajaran; pelajaran

4.2 Lexicography analysis

The definitions and examples in KBBI can be regarded as a corpus which can be employed for various analyses and give further insights to the Indonesian language. We extracted the twenty most frequent words in definitions using the Python NLTK library (http://www.nltk.org) (see Table 7). These frequent words can be used as a part of a lexical set for the Indonesian learner's dictionary we are making now which uses limited words in the definitions and examples. We can also look for the genus words whose result is shown in Table 8. Lim et al. (2016) present the fifty most frequent words and genus words used in definitions in KamusDewan, the authoritative dictionary for Standard Malay. With these data, we can make a comparison of the vocabularies of Indonesian and Standard Malay.

Table 7 Twenty most frequent words in KBBI definitions

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
yang	43,613	untuk	10,312	pada	6,793	dapat	3,020
dan	26,221	dalam	8,638	orang	6,110	tempat	2,970
atau	14,414	di	8,537	tentang	4,746	sebagai	2,917
sebagainya	12,410	tidak	7,756	seperti	3,422	oleh	2,910
dengan	12,016	dari	7,280	ke	3,247	sesuatu	2,851

Table 8 Twenty most frequent genus words in KBBI definitions

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
orang	2,703	tempat	806	keadaan	526	ilmu	401
proses	1,858	hasil	656	ikan	521	fobia	350
alat	1,595	sesuatu	573	hal	512	nama	337
bagian	835	kata	557	tumbuhan	443	zat	300
perihal	823	pohon	547	tiruan	413	penyakit	297

4.3 Linguistic analysis

The Indonesian language has a very rich morphology for word derivation process. It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, non-productive infixes; and a variety of reduplications. Most of the affixes are derivational (Sneddon et al. 2010).Using regular expressions in Python, we made a table of more than 100 patterns of word formation based on affixes and various types of reduplication in Indonesian. Table 9 shows a part of it. It has been used in a linguistics research for analyzing the difference between meN-...-i and meN-...-kan (NurAmirahKhairulAnuar et al. 2017). There are many possibilities we can do with the data, such as analyzing other affixes and reduplications.

Table 9 Some derived words in KBBI, grouped by affixes

Affix/Redup.	Example	Number	Percentage
meN-	mengabadi	5,185	21.1%
meN-...-kan	mengabadikan	2,884	11.7%
ber-	berabang	2,704	11.0%
-an	abaian	1,873	7.6%
peN-...-an	pengabadian	1,780	7.2%
peN-	pengabai	1,552	6.3%

4.4 Linking to other lexical resources

KBBI contains a number of scientific names for flora and fauna. Using them as a pivot, we aligned more than 600 entries in KBBI to the entries in other lexical resources, such as WordnetBahasa (Bond et al. 2014). Table 10 shows some examples of aligned entries via scientific names.

Table 10 Some examples of aligned KBBI entries and the Wordnetsynsets

KBBI entry	Scientific name	Wordnet lemma	Wordnetsynset
abaka	musatextilis	abaca	12353431-n
abalone	haliotis	Haliotis	01942724-n
abrikos	prunusarmeniaca	common apricot	12641007-n
acerang	coleus amboinicus	country borage	12845187-n
adas	foeniculumvulgare	common fennel	12939282-n
adasmanis	pimpinellaanisum	anise, anise plant	12943049-n

4.5 Online and offline applications

KBBI database serves as the vital part in building the online DWS (https://kbbi.kemdikbud.go.id), called 'KBBI DalamJaringan' or KBBI Daring, launched on 28 October 2016 (Kamajaya et al. 2017) and offline mobile applications, both for Android (https://play.google.com/store/apps/details?id=yuku.kbbi5) and iOS (https://itunes.apple.com/us/app/kamus-besar-bahasa-indonesia/id1173573777), launched on 17 November 2016. Figure 15 shows the homepage of the online KBBI and figure 16 shows both the screenshots of the Android and iOS applications. In order to facilitate the workflow of the editorial staff for the online application and the online public participation to add, edit, and deactivate lemmas, definitions, and examples, the KBBI database is equipped with tables for proposals.

Figure 15 Screenshot of the online KBBI homepage

Figure 16 Screenshots of the Android (left) and iOS (right) mobile applications

6. Conclusion and future work

We have described our work in creating a database for KBBI from Microsoft Excel and Word files by converting them to a RTF file and a HTML file, identifying its structure, cleaning up the data, and breaking it down based on the structure. The broken down components were then exported to SQLite database. The database allows lexicographers, linguists, and researchers in NLP field to access the rich lexicographic and linguistic contents in the Indonesian language in more flexible ways, opening up possibilities in discovering new insights into the language, as well as helping the KBBI editorial staff work on the dictionary more effectively.

In the near future, the database will be expanded with etymological information. Our work on compiling and editing the etymological information has been done since 2015 and is still in progress. We have finished working on lemmas from Sanskrit and are working onlemmas originating from Old Javanese and Dutch. In addition, the database will be connected to a corpus. The source of the corpus we are building is from scientific publications. We have finished the first stage and are now adding about five million words per year.

Acknowledgments

Thanks to Francis Bond and Lu ́sMorgado da Costa for the precious advice on the database structure. Thanks to Ivan Lanin for improving the database and making it more efficient. Thanks to Lim LianTze who inspired us to write this paper. Thanks to NTU HSS library support staff: Rashidah Ismail, Raihana Abdul Wahid, and Tan ChuanKo for allowing the first author to borrow the fourth edition of KBBI for months and to Wong Oi May who helped order the dictionary.

References

Alwi, Hasan, SoenjonoDardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 2014. Tata bahasabakubahasa Indonesia [The standard grammar of Indonesian]. Jakarta: BalaiPustaka.

Amalia, Dora (ed). 2016. KamusBesarBahasa Indonesia.5th edition. Jakarta: BadanPengembangandanPembinaanBahasa.

Atkins, B. T. S. and Rundell, M. 2008. The Oxford guide to practical lexicography. Oxford: Oxford University Press.

Bond, Francis, LianTze Lim, Enya Kong Tang, and HammamRiza. 2014. The combined WordnetBahasa. NUSA: Linguistic studies of languages in and around Indonesia 57: 83–100.

Jurafsky, Daniel and James H. Martin. 2009. Speech and language processing. 2nd edition. New Jersey: Pearson Education, Inc.

Kamajaya, Ian, David Moeljadi, and Dora Amalia. 2017. KBBI Daring: arevolution in the Indonesian online dictionary and lexicography. To be presented at eLex 2017. Leiden, September 19-21.

Lewis, M. Paul. 2009. Ethnologue: languages of the world. 16th edition. Dallas: SIL International.

Lim, LianTze, Ruoh Tau Chiew, Enya Kong Tang, Rusli Abdul Ghani, and NaimahYusof. 2016. Digitising a machine-tractable version of KamusDewan with TEI-P5. PeerJ Preprints, 4, e2205v1.

NurAmirahBinteKhairulAnuar, Hannah Choi, and FrantišekKratochvíl. 2017. Verb subcategorization: -kan and -i suffixing verbs in Malay and Indonesian. Presented at The Twenty-First International Symposium On Malay/Indonesian Linguistics (ISMIL 21).Langkawi, May 4-6.

Sneddon, James Neil, Alexander Adelaar, DwiNoveriniDjenar, and Michael C. Ewing. 2010. Indonesian reference grammar. 2nd edition. New South Wales: Allen &Unwin.

Sugono, Dendy (ed.). 2008. KamusBesarBahasa Indonesia PusatBahasa.4th edition. Jakarta: PT GramediaPustakaUtama.