

# Building the *Kamus Besar Bahasa Indonesia* (KBBI) Database and Its Online Application

David **Moeljadi**

Division of Linguistics and Multilingual Studies,  
Nanyang Technological University, Singapore

The 21st International Symposium on Malay/Indonesian Linguistics (ISMIL 21),  
Langkawi Research Center

4 May 2017

1. Kamus Besar Bahasa Indonesia (KBBI)
2. From Word and Excel to Database
3. Features in the Online KBBI V
4. Searching words and making proposals in the Online KBBI V

<sup>1</sup>**ka.mus** *n* **1** **buku acuan** yang memuat **kata** dan ungkapan, biasanya **disusun menurut abjad** berikut keterangan tentang **makna**, pemakaian, atau terjemahannya;...

-- **besar** kamus yang memuat khazanah secara lengkap, termasuk kosakata istilah dari berbagai bidang ilmu yang bersifat umum;...

*Kamus Besar Bahasa Indonesia* Fifth Edition [1]

**dic·tio·nary** *noun* **1** a **reference source** in print or electronic form containing **words** usually **alphabetically arranged** along with information about their forms, pronunciations, functions, etymologies, **meanings**, and syntactic and idiomatic uses

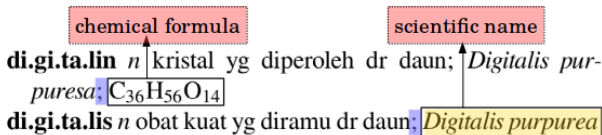
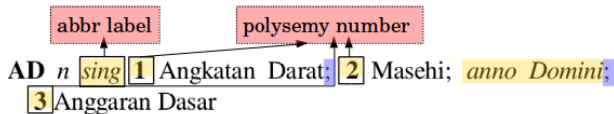
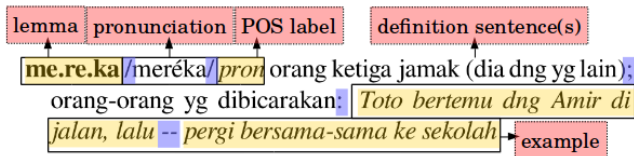
Merriam-Webster

# Kamus Besar Bahasa Indonesia (KBBI)

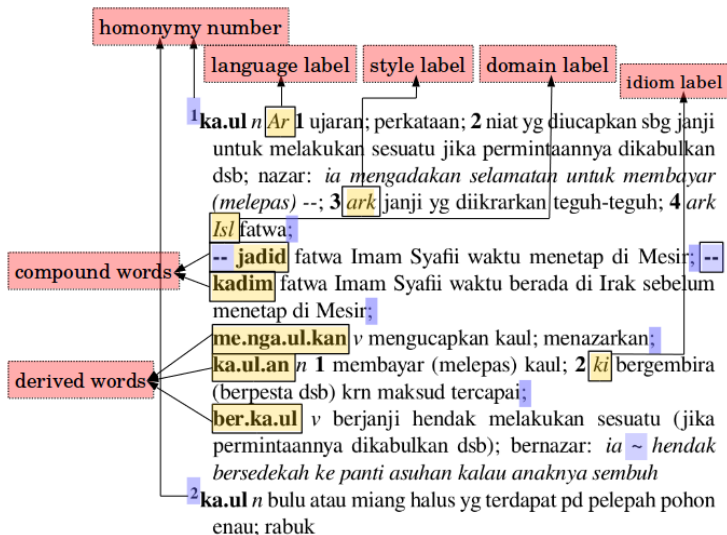


- the official dictionary of the Indonesian language
- published by *Badan Pengembangan dan Pembinaan Bahasa* (The Language Development and Cultivation Agency) or *Badan Bahasa* under Ministry of Education and Culture, Republic of Indonesia
- KBBI Fourth Edition (KBBI IV) [5] had its data in Microsoft Excel and Word files

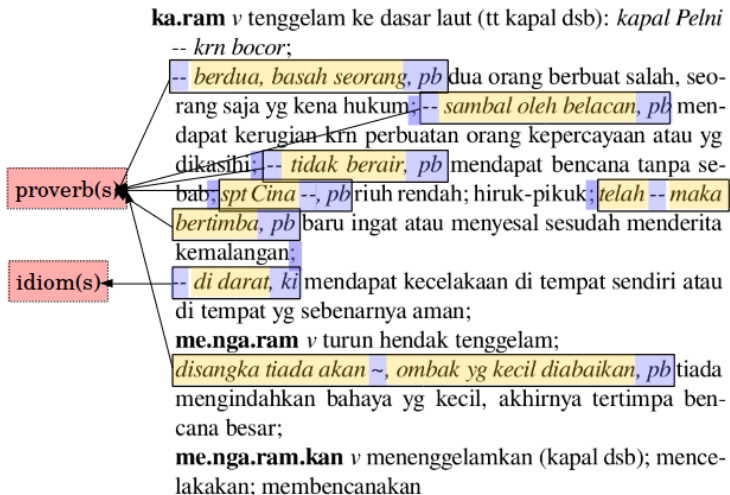
# Dictionary entries in KBBI



# Dictionary entries in KBBI



# Dictionary entries in KBBI



# Dictionary entries in KBBI

## Cross-references

**ke.ron.sang** → **kerongsang**  
**ke.ron.tang** lihat <sup>1</sup>**kering**

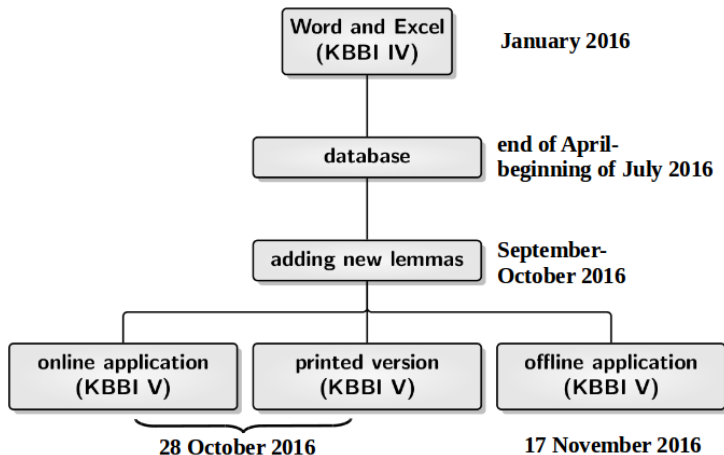


# The Online KBBI before October 2016



- data from KBBI III, for simple word search by root (*kata dasar*)
- the result is exactly in the same format as the one in the printed dictionary
- the data was not structured, no database

# From KBBI IV to KBBI V



## From KBBI IV to KBBI V

[illegible]

## January 2016

end of April-  
beginning of July 2016

**September-  
October 2016**

	id	entri	jenis varian	kolom	status
1	A	dasar	[null]	[null]	
2	A	dasar	n	nama pertama aliyah indonesi	
3	A	dasar	n	nama haruf <u>o>u>	
4	A	dasar	n	penanda pertama dalam urutan (mulus, rifat, ddi)	
5	A	dasar	n	harga tiap tiap satuan	
6	A	dasar	[null]	nama, tahun lahir (nama dua angka untuk memperkirakan panjang, besar, dan sesuai)	
7	A	dasar	n	nama & tahun lahir	



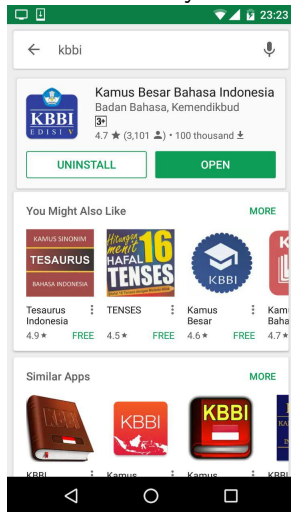
28 October 2016



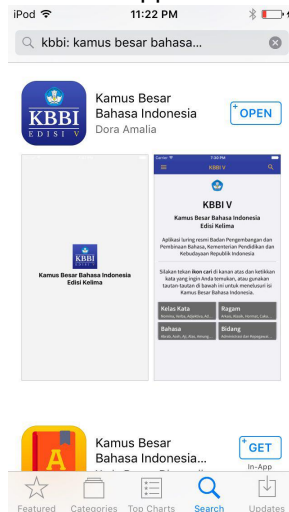
17 November 2016

# Smartphone applications

## Android Play Store



## iOS App Store



Free - gratis!

# Word and Excel files

	A
1	<b>A, a</b> <b>n</b> 1 huruf pertama abjad Indonesia; 2 nama huruf <i>a</i> ; 3 penanda pertama dl urutan (mutu, nilai, dsb)
2	<b>à</b> 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	<b>a-</b> bentuk terikat 1 kekurangan: anemia; 2 tidak atau bukan: aseksual; 3 tanpa: anonim
4	<b>aa</b> <i>Sd n</i> akang
5	<b>1ab</b> <i>n</i> wadah kecil dr timah untuk candu; hap
6	<b>2ab</b> <i>ark n</i> ayah
7	<b>ab-</b> bentuk terikat dari; jauh dr: abnormal
8	<b>aba</b> <i>n</i> ayah; bapak

## A

**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; serampangan;  
**aco-aco.an** *a* ugal-ugalan; sembrono  
**ae.ros.kop** /aéroskop/ *n* alat untuk menangkap debu, bakteri, spora, dsb dr udara untuk tujuan tes (percobaan, pengujian)  
**2agon** *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 terlambat  
**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

# From Word and Excel to Rich Text Format (rtf)

```
\trowd \trgaph30\trleft-30\trrh317\cellx1040\clmgf \cellx2351\clmrg \cellx18546\pard \intbl
\qc \f5\fs22 \cf55 1\cell \ql \f6\fs22 \b A\fs22 \b0 , \f6\fs22 \b a \f7\fs22 \i \b0 n\fs
\fs22 \i0 \f6\fs22 \b 1\fs22 \b0 huruf pertama abjad Indonesia; \f6\fs22 \b 2\fs22
\b0 nama huruf \f7\fs22 \i a\fs22 \i0 ; \f6\fs22 \b 3\fs22 \b0 penanda pertama dl
urutan (mutu, nilai, dsb) \cell \qr \f0\fs22 \cell
\pard \intbl \row\trowd \trgaph30\trleft-30\trrh317\cellx1040\clmgf \cellx2351\clmrg
\cellx18546\pard \intbl \qc \f5\fs22 2\cell \ql \f6\fs22 \b \u224\ 'e0\fs22 \b0 \f6\fs22
\b 1\fs22 \b0 kira-kira; lebih kurang (antara dua angka untuk memperkirakan panjang,
besar, dsb sesuatu): \f7\fs22 \i ular itu panjangnya 6 \u8212\ '97 7 m\fs22 \i0 ; \f7\fs22
\i lama perjalanan 2 \u8212\ '97 3 jam\fs22 \i0 ; \f6\fs22 \b 2\fs22 \b0 harga tiap-
tiap satuan: \f7\fs22 \i ia membeli bahan itu 5 m \u8212\ '97 Rp20.000,00\fs22 \i0 \cell
\qr \f0\fs22 \cell
```

# From rtf to HyperText Markup Language (html)

**A**, **a** <i>n</i> **1** huruf pertama abjad Indonesia; **2** nama huruf <i>a</i>; **3** penanda pertama dl urutan (mutu, nilai, dsb)

**à** **1** 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>; **2** harga tiap-tiap satuan: <i>ia membeli bahan itu 5·m \u8212'97·Rp20.000,00</i>

**a-** <i>bentuk terikat</i> **1** kekurangan: <i>anemia</i>; **2** tidak atau bukan: <i>aseksual</i>; **3** tanpa: <i>anonim</i>

**aa** <i>Sd</i> <i>n</i> akang

**ab (1)** <i>n</i> wadah kecil dr timah untuk candu; hap

**ab (2)** <i>ark</i> <i>n</i> ayah

**ab-** <i>bentuk terikat</i> dari; jauh dr: <i>abnormal</i>

**aba** <i>n</i> ayah; bapak

# Using Python...

```
for line in f.readlines():
    try:
        items = line.strip()
        #####
        ##### ENTRY WORDS #####
        #####
        # First table for entry words
        # search the entry words
        if bool(re.search(r'^<b>', items)):
            # extract the entry words
            lemma = re.findall(r'^<b>(.*?)</b>', items)
            word.append(lemma[0])
            master.append(lemma[0])
            ##### FOR WORD VARIANTS #####
            # change "alf(u)" to "alf" (variant 1) and "alfu" (variant 2)
            if bool(re.search(r'^S(\D+)', lemma[0])):
                lemma_var = [(re.findall(r'(.+?)\(\D+\)', lemma[0])[0], re.findall(r'(.+?)
\(\D+\)', lemma[0])[0]+re.findall(r'(.+?)\(\D+\)', lemma[0])[0])]
                allomorph.append(lemma_var[0][1])
                master.append(lemma_var[0][1])
                lemma_var_without_dots = re.sub(r'\.', '', lemma_var[0][1])
                outputLine = str(len(allomorph)) + '\t' + str(len(word)) + '\t' + ' ' + '\t'
+ lemma_var_without_dots
                outputVar.append(outputLine)
                masterLine = str(len(master)-1) + '\t' + lemma_var_without_dots + '\t' +
'variant' + '\t' + str(len(allomorph))
                outputMaster.append(masterLine)
            # change "-anda (-nda, -da)" to "-anda" (variant 1) and "-nda, -da" (variant 2)
            elif bool(re.search(r'^s+(\D+)', lemma[0])):
```

The data was broken down by **lemmas**, **sublemmas** (derived words, compounds, proverbs, and idioms), **labels**, **pronunciations**, **definitions**, **examples**, **scientific names**, and **chemical formulas** using **regular expression**, 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 [4].



# Regular expression

clean.txt x

```
1 <b>A</b> , <b>a</b><i>n</i> <b>l</b> huruf pe  
huruf<i>a</i>; <b>3</b> penanda pertama dl ur  
2 <b>a</b> <b>l</b> kira-kira; lebih kurang (an  
panjang, besar, dsb sesuatu):<i>ular itu panja  
perjalanan 2 -- 3 jam</i>; <b>2</b> harga tia  
itu 5 m -- Rp20.000,00</i>  
3 <b>a-</b> <i>bentuk terikat</i><b>l</b> kekura  
atau bukan:<i>aseksual</i>; <b>3</b> tanpa:<i>  
4 <b>aa</b><i>Sd</i><i>n</i> akang  
5 <b>ab (1)</b> <i>n</i> wadah kecil dr timah u  
6 <b>ab (2)</b> <i>ark n</i>ayah
```

Replace

Search for:

`^<b>.+?</b>`

Replace with:

☐ Match case

☐ Match entire word only

☒ Match as regular expression

☐ Search backwards

☒ Wrap around

Close

Replace All

Replace

Find

`^<b>.+?</b>`

`^` : searches at the start of each line

`<b>.+?</b>` : searches everything  
between `<b>` and `</b>`

## SQLite ([www.sqlite.org](http://www.sqlite.org))

Database	
▼	main
▼	Tables (17)
▶	bahasa
▶	berimbuhan
▶	bidang
▶	contoh
▶	entri
▶	gabungan
▶	idiom
▶	ilmiah
▶	kata
▶	kelaskata
▶	kimia
▶	makna
▶	maknacontoh
▶	peribahasa
▶	ragam
▶	rujuk
▶	varian
▶	Views (0)

eid	entri	jenis	kelas	makna
1	a	varian	{null}	{null}
2	A	dasar	n	huruf pertama abjad Indonesia
2	A	dasar	n	nama huruf <i>a</i>
2	A	dasar	n	penanda pertama dalam urutan (mutu, nilai, dsb)
3	à	dasar	{null}	harga tiap-tiap satuan
3	à	dasar	{null}	kira-kira; lebih kurang (antara dua angka untuk memperkirakan panjang)
4	a-	dasar	bentuk terikat	kekurangan

# The current state of the KBBI Database

- Lemmas: 48,141
- Derived words: 26,197
- Compound words: 30,376
- Proverbs: 2,040
- Idioms: 267
- Entries (total): 108,241
- Definition sentences: 126,639
- Examples: 29,254

# What can we get from KBBI Database? I

## ① More specific and targeted word lookups, e.g.

- ▶ looking up **phrases and MWEs** such as compound words, idioms, and proverbs as well as **derived words**

```
SELECT entri, jenis, makna FROM baseview WHERE entri="sedia payung sebelum hujan";
```

	entri	jenis	makna
1	sedia payung sebelum hujan	peribahasa	bersiap sedia sebelum terjadi yg kurang baik

- ▶ looking up entries by their **labels** (part-of-speech, language, and domain labels)

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

	entri	ragam	bahasa	makna
1	cutel	ark	Jw	tamat; habis (tt cerita dsb); berakhir
2	gundang	ark	Jw	lekum; tenggorok
3	pembarap	ark	Jw	anak sulung
4	sikep	ark	Jw	orang dr desa yg mempunyai kewajiban melakukan kerja
5	ubel-ubel	ark	Jw	tentara Inggris asal India
6	wiyata	ark	Jw	pengajaran; pelajaran

# What can we get from KBBI Database? II

## ② Lexicography analysis

- ▶ extracting the **most frequent words** in the definition sentences → can be used as a **lexical set** for the **Indonesian learner's dictionary**

Word	Freq.	Word	Freq.	Word	Freq.
yang	43,613	untuk	10,312	pada	6,793
dan	26,221	dalam	8,638	orang	6,110
atau	14,414	di	8,537	tentang	4,746
sebagainya	12,410	tidak	7,756	seperti	3,422
dengan	12,016	dari	7,280	...	...

- ▶ extracting the **most frequent genus terms** in the definition sentences

Word	Freq.	Word	Freq.	Word	Freq.
orang	2,703	perihal	823	sesuatu	573
proses	1,858	tempat	806	kata	557
alat	1,595	menjadikan	745	pohon	547
tidak	1,526	yang	664	mempunyai	526
bagian	835	hasil	656	...	...

# What can we get from KBBI Database? III

## ③ Linguistic analysis

- ▶ grouping the derived words based on **affixes** and patterns of **reduplication** in Indonesian

Affix/Redup.	Example	Number	Percentage
meN-	<b>mengabadi</b>	5,185	21.1%
meN-...-kan	<b>mengabadikan</b>	2,884	11.7%
ber-	<b>berabang</b>	2,704	11.0%
-an	<b>abaian</b>	1,873	7.6%
peN-...-an	<b>pengabadian</b>	1,780	7.2%
...	...	...	...
<b>Total</b>		24,587	100.0%

## ④ Online and offline applications etc.

# The Online KBBI V



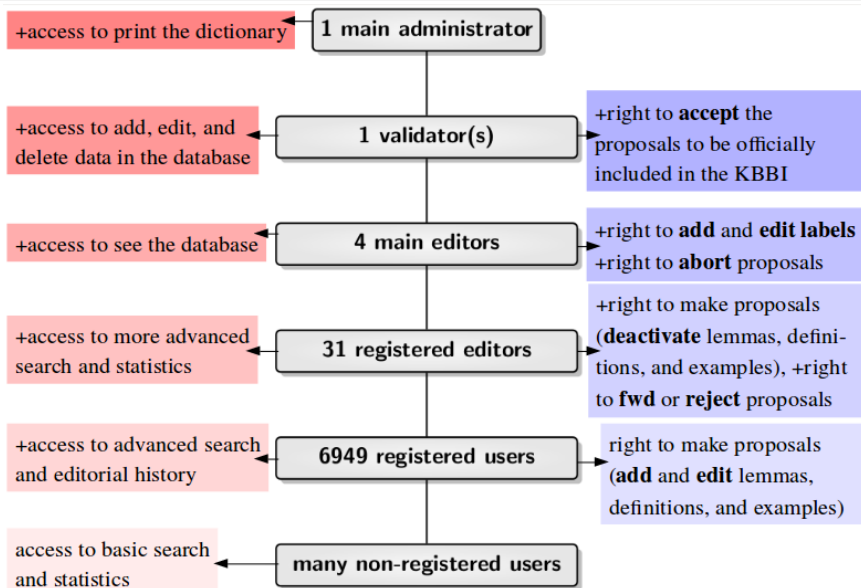
- officially launched on 28 October 2016 [1], its user interface and the system were made using ASP.NET ([www.asp.net](http://www.asp.net)).
- <https://kbbi.kemdikbud.go.id/>
- **Dictionary Writing System (DWS)** [2] which enables lexicographers to compile and edit dictionary text, as well as to facilitate project management, typesetting, and output to printed or electronic media

## Some features in the Online KBBI

	Before 28 Oct 2016	After 28 Oct 2016
<b>Word search</b>	basic (by roots)	advanced (+by labels etc.)
<b>Lexicographical workflow</b>	done within the editorial board in Badan Bahasa	+online public participation to add, edit, and deactivate lemmas, definitions, and examples (crowdsourcing)
<b>Data format</b>	several inconsistencies	more consistency
<b>Security system</b>	data can be easily crawled	customized security system to protect the data from web crawlers
<b>Print function</b>	no print function	print function can convert the data in the database to print format



# Lexicographical workflow in the Online KBBI



# How a new lemma can be included in KBBI?

- 1 Having a unique concept  
**NOT OK** **si.ha.lu.an** *v* saling bertemu (cf. **ber.se.mu.ka**)
- 2 According to the Indonesian spelling rules  
**NOT OK** **ojeg** *n* sepeda atau sepeda motor yang ditambahkan dengan cara memboncengkan penumpang atau penyewanya (cf. **ojek**)
- 3 Euphonic (being pleasing to the ear)  
**NOT OK** **la.bu.la.bu.wai** *n* nasi yang diberi air putih ditambah garam atau ikan asin
- 4 Having positive connotations
- 5 Having a high frequency of use and a broad range of users




Dora Amalia, p.c.

# Rejected proposal

## Usulan (Pid: 8125, Pengusul: David Moeljadi)

[Tutup Usulan](#)[Tutup Detail](#)

### Keterangan

Pid	8125
Entri	LU
Jenis	 Ubah
Elemen	 Makna
Status	 Ditolak
Eid	96073
Id Tabel	112999
Induk Pid	(Tidak tersedia)
Anak Pid	(Tidak tersedia)

change  
definition  
rejected

### Pembuat

Nama	David Moeljadi
Pos-el	<a href="mailto:davidmoeljadi@gmail.com">davidmoeljadi@gmail.com</a>
Tingkat	Editor
Dibuat	2017-02-06 14:22:23.645
Dikirim	2017-02-06 14:22:23.833
Penjelasan	saya ubah "lintang utara" menjadi "Lintang Utara" karena BB bermakna "Bujur Barat" dan BT "Bujur Timur" (keduanya ditulis berawalan huruf kapital).


I changed "lintang utara" to "Lintang Utara" because BB's definition is "Bujur Barat" and BT's definition "Bujur Timur" (both starts with capital letters)

### Editor

Nama	(Tidak tersedia)
Pos-el	(Tidak tersedia)
Aksi	(Tidak tersedia)
Diproses	(Tidak tersedia)
Penjelasan	(Tidak tersedia)

registered editor

### Redaktur

Nama	Adi Budiwiyanto
Pos-el	<a href="mailto:adi.budiwiyanto@kemdikbud.go.id">adi.budiwiyanto@kemdikbud.go.id</a>
Aksi	 Ditolak
Diproses	2017-03-18 22:34:04.078
Penjelasan	yang ini sudah benar; silakan usulkan perbaikan untuk BB dan BT

main editor

rejected

### Validator

Nama	(Tidak tersedia)
Pos-el	(Tidak tersedia)
Aksi	(Tidak tersedia)
Diproses	(Tidak tersedia)
Penjelasan	(Tidak tersedia)

this one is correct; please make new proposals for BB and BT




# Accepted proposal

Usulan (Pid: 8022, Pengusul: David Moeljadi)

Tutup Usulan

Tutup Detail

## Keterangan

Pid	8022
Entri	berpotongan
Jenis	 Ubah
Elemen	 Makna
Status	 Diterima
Eid	64851
Id Tabel	128258
Induk Pid	6756
Anak Pid	(Tidak tersedia)

change  
definition  
accepted

## Pembuat

Nama	David Moeljadi
Pos-el	<a href="mailto:davidmoeljadi@gmail.com">davidmoeljadi@gmail.com</a>
Tingkat	Editor
Dibuat	2017-02-05 12:00:10.719
Dikirim	2017-02-06 14:15:46.181
Penjelasan	<i>ditambahkan informasi kelas kata (verba) dan makna "saling memotong"</i>


added POS label (verb) and  
a definition "saling memotong"

## Editor

Nama	(Tidak tersedia)
Pos-el	(Tidak tersedia)
Aksi	(Tidak tersedia)
Diproses	(Tidak tersedia)
Penjelasan	(Tidak tersedia)

registered editor

## Redaktur

Nama	Menuk Hardaniwati
Pos-el	<a href="mailto:menuk.hardaniwati@kemdikbud.go.id">menuk.hardaniwati@kemdikbud.go.id</a>
Aksi	 Diterima
Diproses	2017-02-06 14:42:53.845
Penjelasan	(Tidak tersedia)

main editor

accepted

## Validator

Nama	Dora Amalia
Pos-el	<a href="mailto:dora.amalia@kemdikbud.go.id">dora.amalia@kemdikbud.go.id</a>
Aksi	 Diterima
Diproses	2017-02-09 13:06:57.806
Penjelasan	ok

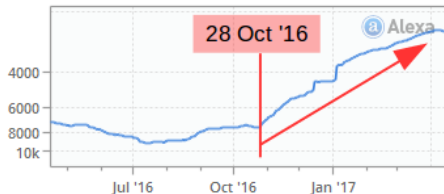
ok

# Current situation (as of 4 May 2017 10:10 am)

- Word lookups
  - ▶ Total: 3,015,927 (11.16/minute, 669.37/hour, 16,065.00/day)
- Proposals
  - ▶ Total: 9,269 (49.37/day)
  - ▶ Accepted: 2,720
  - ▶ Rejected: 501
  - ▶ Being processed: 5,571
- Popularity (according to Alexa Traffic Ranks [www.alexa.com](http://www.alexa.com))
  - ▶ Global rank: 2,695
  - ▶ Rank in Indonesia: 66

## Alexa Traffic Ranks

How is this site ranked relative to other sites?



# Future work

- add etymological information
- connect to corpora
- link to other lexical resources such as Wordnet Bahasa [3]

# (1) Searching words in the Online KBBI

- ➊ Go to <https://kbbi.kemdikbud.go.id>
- ➋ Register as a new user
- ➌ Check your email inbox
- ➍ Click the link in the email
- ➎ Search words by:
  - ▶ root words
  - ▶ orthography
  - ▶ labels: parts-of-speech, language, domain, style, type

## (2) Making proposals in the Online KBBI

- Add new words
- Edit some definitions
- Add new examples



# Acknowledgments

- Thanks to Dora Amalia for the KBBI IV data and her support
- Thanks to Francis Bond and Luis Morgado da Costa for the precious advice on the database structure
- Thanks to Ivan Lanin for improving the database
- Thanks to Ian Kamajaya for building the Online KBBI
- Thanks to Randy Sugianto for creating the Android application
- Thanks to Jaya Satrio Hendrick for designing the Android and iOS applications
- Thanks to Lie Gunawan for creating the iOS application
- Thanks to NTU HSS library support staff: Rashidah Ismail, Raihana Abdul Wahid, and Tan Chuan Ko for allowing me to borrow KBBI IV paper dictionary for months; and to Wong Oi May who helped us order the dictionary

# References

Dora Amalia, ed. *Kamus Besar Bahasa Indonesia*. 5th ed. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, 2016.

B. T. Sue Atkins and Michael Rundell. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008.

Francis Bond et al. “The combined Wordnet Bahasa”. In: *NUSA: Linguistic studies of languages in and around Indonesia* 57 (2014), pp. 83–100.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. 2nd ed. New Jersey: Pearson Education, Inc., 2009.

Dendy Sugono, ed. *Kamus Besar Bahasa Indonesia Pusat Bahasa*. 4th ed. Jakarta: PT Gramedia Pustaka Utama, 2008.

**te.ri.ma ka.sih** *n* rasa syukur;

**ber.te.ri.ma ka.sih** *v* mengucapkan syukur; melahirkan rasa syukur atau membalas budi setelah menerima kebaikan dsb