# CONSIDERATIONS FOR PROVIDING ETYMOLOGICAL INFORMATION IN THE KBBI INDONESIAN DICTIONARY

**David Moeljadi**

**Palacký University Olomouc**

**Ian Kamajaya**

**ASTrio Pte Ltd**

**Azhari Dasman Darnis**

**Badan Bahasa**

## Abstract

We discuss the inclusion of etymological information in the Indonesian dictionary KBBI (Kamus Besar Bahasa Indonesia) fifth edition. KBBI is the most comprehensive and authoritative Indonesian monolingual dictionary, published/launched by The Language Development and Cultivation Agency, under the Ministry of Education and Culture (Moeljadi et al. 2017, Kamajaya et al. 2017). It is mainly online-based (https://kbbi.kemdikbud.go.id), updated regularly, and will be enriched by etymological information from October 2019. This etymological information is valuable for the Indonesian language that has loanwords from various languages and language families: Austronesian (Old Javanese), Indo-European (Sanskrit, Persian, Portuguese, Dutch, English), Dravidian (Tamil), Semitic (Arabic), and Sinitic (Hokkien, Cantonese, Mandarin). The first etymology project began in the late 2000s, after the fourth edition was published. A team was formed based on the grouping of donor languages: (1) Arabic and Semitic languages, including Persian; (2) European languages (Dutch, English, and French); (3) Old Javanese; and (4) Chinese languages. Unfortunately, this project did not meet the target. Starting from the second etymology project, since 2016 each year we focus on two groups and involve experts from universities in Indonesia. We refer to previous work and references e.g. Jones et al. (2007). Etymology projects on Sanskrit and Old Javanese were carried out (2016~2017), after that Dutch (2017~) and Arabic (2018~). Data collection is based on the dictionary headwords. The technical part involves programming and database restructuration. The existing KBBI database will be augmented with etymological information-related tables containing the original scripts of the loanwords and the relationships within them as well as between them and the entries, enabling KBBI to present etymological relations of the loanwords accurately. We believe that the etymological information in KBBI will serve as a valuable resource and accompaniment to Indonesian historical, linguistics, and lexicography research.

**Key Words:** Indonesian, etymology, online dictionary, database structure

## Introduction

Kamus Besar Bahasa Indonesia (KBBI) is the official dictionary of the Indonesian language, published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency) or Badan Bahasa, under the Ministry of Education and Culture, Republic of Indonesia. Up until present, KBBI is the most comprehensive and the most authoritative reference for the Indonesian language. Its first edition, 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). Its third edition, published in 2001, contains 78,000 entries and seven years later, the number of entries in the fourth edition increased to more than 92,000. Its latest, fifth edition was released for the first time in 2016 in three formats: printed, online, and offline mobile versions. Since then, it is regularly updated twice a year. As of April 2019, it has more than 110,000 entries. Moeljadi et al. (2017) describes the creation of the database as well as the database structure. Kamajaya et al. (2017) explains the online KBBI in details. Although KBBI has comprehensive data on headwords, derived words, compounds, proverbs, word senses, parts-of-speech etc., it does not have information on the word origins or etymology: from which language(s) a particular word is borrowed, the etymological route it takes until it enters the Indonesian language, as well as changes in word forms, sounds, and meanings. This paper discusses the inclusion of this etymological information into KBBI database, especially regarding the database structure.

## Historical background and contact with other languages

For centuries the ancestors of Indonesian speakers had a contact with speakers from various nations in the world. Sanskrit is recorded as the first foreign language which entered Indonesia, since the beginning of A.D. This language became the language of literature and also a medium of spread of Hinduism and Buddhism. Hinduism spread vastly in Java in the seventh and eighth century, then Buddhism in the eighth and ninth century. Together with the spreading of Hinduism, the spice trade with Indians also took place. Some of the Indians are Hindi speakers, while some are Tamil speakers from Southern India and Eastern Sri Lanka, whose language became the medium of literary work. Tamil language had a strong influence on Malay language.

Contact with Chinese speakers had happened since the seventh century, when Chinese merchants traded to Riau Islands, West Kalimantan, and East Kalimantan, even until North Maluku. When Sriwijaya kingdom appeared and became strong, China also opened a diplomatic relation with Sriwijaya to secure its trade and shipping business. In the year 922, Chinese travelers visited Kahuripan kingdom in East Java. Since the 11th century, hundreds of thousands of Chinese migrants left their ancestral land and settled in many parts of the Archipelago. What is called

“Chinese” here, actually is more accurately said as languages from China. There are many languages in China. Four of them are well-known languages in Indonesia, i.e. Hokkien, Hakka, Cantonese, and Mandarin. Because the contact had lasted long, it is reasonable that many loanwords from Chinese languages came into Malay/Indonesian. However, because the Chinese languages were not used as religious, scientific, or literary medium in Indonesia, it is very likely that many of those loanwords blended into languages in Indonesia.

The Arabic language was brought to Indonesia from the seventh century by merchants from Persia, India, and Arab who were also the spreaders of Islam. Arabic, an Islamic religious language, began to influence Malay, especially since the twelfth century when many kings embraced Islam. Because many of those merchants were Persian speakers, quite a number of Persian words entered into Malay.

The Portuguese language had begun to be known by Malay-speaking community since the Portuguese occupied Malacca in 1511 after they occupied Goa one year before. Portuguese were unable to compete with the Dutch who came later. They stepped aside to the eastern region of the Archipelago. However, in 17th century Portuguese, together with Malay, had become a lingua franca among ethnic groups in the Archipelago.

The Dutch people started to come to the Archipelago in the beginning of 17th century when they expelled Portugues from the Moluccas (Maluku) in 1606 and then headed to the Java island and other regions in the west. Since then, gradually the Dutch people took control of many regions in Indonesia. The Dutch language could not completely oust the Portuguese language because Dutch was more difficult to learn and Dutch people did not like to open up to people who wanted to learn Dutch culture, including the language. However, the Dutch occupation was gradually covering almost the entire country and lasted for a long time. The Netherlands was also a main source of learning for youth groups in the independence movement. Because of that, the concept of state building mostly refers to the Dutch language.

The British once occupied Indonesia although not for long. Raffles invaded Batavia (now Jakarta, the capital city of Indonesia) in 1811 and stayed there for five years. Before he was moved to Singapore, he also stayed in Bengkulu in 1818. Actually, in 1696 the British once sent a messenger Ralph Orp to Padang but landed in Bengkulu and settled there. In Bengkulu the British built a fort named Fort Marlborough in 1714-1719. It means that more or less a contact with English had already happened for a long time in a region near the center of Malay-speaking community. However, the most intensive contact with English happens in this globalization era.

Several years before the independence in 1945, Japan occupied Indonesia. However, the Japanese occupation lasted only for three and a half years and it left very few words which can survive through generations.

Because of this historical background and the richness of loanwords in Indonesian, many etymological work have been done by various researchers, such as Jones (2007) and Tadmor (2009). In Indonesia, some small-scale research has been done before the year 2000, such as *Kamus Etimologi* “Etymological Dictionary” (Harun et al. 1984) that was originally a research report compiled by a research team, it contains only words borrowed from Arabic; *Kamus Etimologi Bahasa Indonesia* “Indonesian Etymological Dictionary” (Adiwimarta et al. 1987), one of the results of Jakarta project research in 1984-1985 which was done by one team of researchers; and *Senarai Kata Serapan dalam Bahasa Indonesia* “A List of Loanwords in Indonesian” (Jumariam et al. 1996). However, to the best of our knowledge, there is no comprehensive work for Indonesian loanwords that has an open-source database that can be used to augment the existing KBBI database.

## Method and Results

The idea of the inclusion of etymological information in KBBI was proposed to the KBBI's editorial board in the early 2000s. This is in line with one of the recommendations from the Congress of the Indonesian Language in 2003, i.e. to provide etymological information in KBBI, since the information of the word origins in KBBI is very simple and limited. KBBI provides only information about the donor language for some headwords, for example the word "kamsia" is labeled *Cn* (Chinese), without its original form, sound, and meaning in Chinese.[12]

In 2010 the first etymological project of KBBI started. The main purpose is to provide KBBI with reliable etymological information, i.e. the origin of the words, the changes both in the sense and form, and the etymological route each word took before it was borrowed into Indonesian, e.g.

> **butik** *n* toko tempat menjual pakaian jadi dengan segala kelengkapannya (terutama untuk wanita)
>
> [< Bld *boutiek* 'toko kecil, eksklusif, menjual barang/mode mewah; rumah mode' < Pr *boutique* 'bagian depan rumah atau gedung tempat penjual, perajin memamerkan dan menjajakan, serta menjual dagangannya; warung, kedai, toko' < Yn *apothēkē* 'gudang penyimpanan sediaan']

In the beginning, the project involved lexicographers from Pusat Bahasa.[13] None of them had the expertise in etymology. They relied on dictionaries and etymological research reports. One of the

[12] https://kbbi.kemdikbud.go.id/entri/kamsia

[13] Pusat Bahasa is the former name of Badan Bahasa (The Language Development and Cultivation Agency).

reports is from the research conducted by Pusat Bahasa under the title *Senarai Kata Serapan dalam Bahasa Indonesia* (SKSBI) “A List of Loanwords in Indonesian” (Jumariam et al. 1996). Another one is from an Indonesian Etymological Project conducted by KITLV, Leiden, summarized in a book “Loan-Words in Indonesian and Malay” (LWIM) (Jones 2007). The latter becomes the standard of how the etymological information will be presented in KBBI.

The project began with the list of entries in SKSBI and LWIM. SKSBI lists words in KBBI borrowed from 11 languages: Arabic, Dutch, Chinese, English, Portuguese, Sanskrit, Old Javanese (OJ), Hindi, Persian, Tamil, and Malay. Entries were listed separately based on the donor languages and compared with the ones in LWIM and with dictionaries of donor languages. The words of English origin, for example, were verified using Merriam-Webster Dictionary, Oxford Dictionary, Chambers Dictionary of Etymology, Dictionary of English Etymology, and Microsoft Encarta 2009. The lemma *carter* in SKSBI (page 96) is from English “charter”, its meaning is “the hiring of something for a special purpose”. In LWIM, it is written “**carter** [charter; rent, let] < Eng *charter*”. Afterwards, it is compared with other sources to check the original form and meaning, as well as the etymological route it took before entering Indonesian.

> **carter** *v* [< Ing *charter* 'menyewa kendaraan untuk tujuan pribadi atau khusus' < Pr *chartre* < Lt *chartula* < *charta* 'kertas, daun papirus']

However, because of some budget and priority issues, in 2012 the project was pushed aside and had to take a break for a year. The result of this two-year work is 398 lemmas with etymological information. Most of them have the information only on original forms and senses, without etymological routes. After a one-year break, the project started again.

In 2013 some changes and improvements were made. Some experts were involved. Experts of Arabic, Dutch, Chinese, and Persian from several universities were asked to verify the work done in the previous project as well as to complete the unfinished ones. The project run until 2015 because in 2015 a project of building a KBBI database and online KBBI was given more priority. The result of this four-year work is shown in Table 1.

**Table 1: Result of KBBI etymological project 2010-2012, 2013-2015**

| | **Arabic** | **Dutch** | **English** | **Chinese** | **OJ** | **Korean** | **Sanskrit** | **Persian** | **Japanese** |
|---|---|---|---|---|---|---|---|---|---|
| Alphabet | a, i, j, k, l, m, n, r, s, t | a, f, h, i, l, m, p, q, r, s | c, d, e, f, g, h, I, j, k, p, r, s | all | a--y | all | all | all | all |
| Not-edited | 500 | 846 | 824 | - | 1156 | 64 | 76 | 31 | 275 |
| Edited | 850 | - | - | 432 | - | - | - | - | - |
| Verified | - | - | - | - | - | - | - | - | - |

The etymological information for loanwords from Old Javanese, Korean, Sanskrit, Persian, and Japanese is limited. It does not include the etymological routes. Only Chinese and some of Arabic data was edited.

The etymology project that stopped in 2015 resumed a year later in 2016. It focused on: (1) editing and completion of etymological information of loanwords from Arabic, Sanskrit, Old Javanese, and Dutch; (2) verification of etymological information of loanwords from Chinese, Sanskrit, Old Javanese, and Arabic; and (3) preparation to build a database system. To complete these objectives, a number of experts was involved. Language experts focused on the completion, editing, and verification of etymological information. Information technology experts focused on building a database system. In order to build the database, the data has to be written in a specific format. The online KBBI is planned to be updated in October 2019 with etymological information of loanwords from Arabic. In 2020 it will be updated with information on loanwords from Chinese, Dutch, Sanskrit, and Old Javanese. Table 2 shows the result of work from 2016.

**Table 2: Result of KBBI etymological project 2016-present (April 2019)**

| | **Arabic** | **Dutch** | **English** | **Chinese** | **OJ** | **Korean** | **Sanskrit** | **Persian** | **Japanese** |
|---|---|---|---|---|---|---|---|---|---|
| Alphabet | a, i, j, k, l, m, n, r, s, t | a, f, h, i, l, m, p, q, r, s | c, d, e, f, g, h, I, j, k, p, r, s | all | a--y | all | all | all | all |
| Not-edited | - | - | 824 | - | - | 64 | - | 31 | 275 |
| Edited | 1450 | 1938 | - | - | 303 | - | 932 | - | - |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Verified | - | - | - | 432 | - | - | - | - | - |

The work on loanwords from Arabic, Dutch, Old Javanese, and Sanskrit has arrived at the editing stage. Loanwords from Chinese have been edited and the verification process is in progress. Loanwords from English, Korean, Persian, and Japanese are going to be edited and verified next year due to scale of priorities and shortage of budget. Some words which are regarded as of Old Javanese origin in Table 1, were edited and regarded as of Sanskrit origin in Table 2. Old Javanese and Sanskrit experts distinguished loanwords from Old Javanese and Sanskrit, especially the ones which have similar forms and senses. The project is still continuing until all loanwords are verified. It is planned to end in 2020. At the final stage, the focus is on the data input of the etymological information into KBBI. The project will result in 5,012 loanwords with complete etymological information to enrich the lemmas in KBBI.

## Discussion

This section discusses the KBBI database restructuration/augmentation for etymological information. We build three additional tables as follows:

1. Etymological Source Table, consists of a list of sources/references for etymological information.
2. Etymon Table, consists of a list of etymons together with the information on donor languages, original senses, linguistic information (e.g. part-of-speech, person, number, gender, tense, aspect, voice), and sources/references.
3. Entity Table, represents a ‘minimally-designed’, ‘properties-augmented’ abstract object called

‘entity’ which we build mainly to represent complex logical-etymological relationships between etymons and headwords in KBBI.

Each table is discussed in the next subsections.

### Etymological Source Table

This table consists of a list of etymological sources, their identifiers (IDs), and notes. At present, it has three columns as follows:

1. Etymological Source ID, consists of one unique ID for each source
2. Etymological Source, consists of source names (books, dictionaries, reports etc), author names, years, publishers
3. Notes, consists of optional notes or descriptions for each source

### Etymon Table

Each row in this table basically consists of a set of an etymon, its donor language, and its original sense. Each set has a unique identifier. It is possible that two or more etymons have exactly the same form but come from different languages of origin or they are homonyms, having different senses. The unique identifier will resolve the ambiguity. This table consists of 14 columns as follows:

1. Etymon ID, consists of one unique ID for each set of etymon, donor language, and sense
2. Etymon (original script), consists of a word or morpheme written in the original script, from which a later word is derived
3. Etymon (transliteration), consists of a word or morpheme transliterated or written in Latin

script if the original script is not in Latin script

4. Donor Language, consists of the language of origin
5. Sense, consists of the original meaning of the etymon
6. Etymological Source ID, consists of a set of etymological source IDs. These IDs are linked to the ones in the etymological source table.
7. Part-of-speech, consists of the part-of-speech (noun, verb, adjective etc.) of the etymon
8. Person, consists of information on first, second, and third person, if relevant
9. Number, consists of information on number, such as singular, dual, and plural, if relevant
10. Gender, consists of information on gender, such as masculine, feminine, and neutral, if relevant
11. Tense, consists of information on tense, such as past, present, and future, if relevant
12. Aspect, consists of information on aspect, such as perfect, imperfective, and progressive, if relevant
13. Voice, consists of information on voice, such as active, passive, and medial, if relevant
14. Case, consists of information on case, such as nominative, genitive, dative, and accusative, if relevant

**Entity Table**

The last, and perhaps the most complicated, step we do is to build the 'entity' table. We need to build a 'building-block unit' to represent etymological relationships between one etymon and another etymon, as well as between one etymon and one headword. We use a certain design philosophy, i.e. to build 'minimum' building-block units which are capable of encompassing all the 'logical-etymological' relationships. We call this minimum building-block unit 'entity'. Therefore, an 'entity' is a 'minimally-designed', 'properties-augmented', abstract object, consisting of one reference of etymon and logical-etymological relation to two entities and/or one headword. This table consists of 7 columns, as follows:

1. Entity ID, consists of one unique ID of entity for each row
2. Headword ID, consists of headword ID in the existing KBBI database that is directly related to the respective entity. This ID serves as the primary link to the current KBBI database.
3. First reference entity ID, consists of the reference ID of an entity that is the base of the respective entity.
4. Second reference entity ID, consists of the reference ID of another entity that is the base of the respective entity.
5. Logic, consists of the logical-etymological relationship between the respective entity and its 'referenced entities'. At present, there are five possible values: ROOT, AND, OR, A-AND, and A-OR.
    a. Logic ROOT makes use of the value only in the first reference entity ID column (the second column is left blank) and is used if the first reference entity is the root for the respective entity.
    b. Logic AND is used if the first and the second reference entities are combined to form the respective entity.
    c. Logic OR is used if either of the reference entities is a possible root for the respective entity.
    d. Logic A-AND, stands for 'Augmented-AND', is used to augment one more AND logic in the same level between two reference entities, one of which must already have AND or A-AND logic. This is essentially used to create three or more same-level AND relationships between entities, which would not be possible to be represented by having only two entity reference columns. Note that although logically (1) (X AND Y) AND Z, (2) X AND (Y

AND Z), and (3) X AND Y AND Z have the same result, they are not etymologically identical. For case (1), X AND Y form a word P and then P AND Z form a word W. For case (2), Y AND Z form a word Q and then X AND Q form a word W. For case (3), X AND Y AND Z together form a word W. Adding another A-AND in reference to an entity which already has A-AND will result in augmented AND in the same possible lowest level. Suppose an entity W is formed from X A-AND (Y AND Z), i.e. W = X AND Y AND Z, referencing an entity V with the entity W by using A-AND will result in V A-AND (X AND Y AND Z), i.e. V AND X AND Y AND Z. It can be easily seen that this A-AND logic allows augmentation of as many AND logic in the same level as possible.

e. Logic A-OR, stands for 'Augmented-OR', is used to augment one more OR logic in the same level between two reference entities, one of which must already have OR or A-OR logic. The use of this logic is the same as its A-AND counterpart, i.e. to create a same-level OR relationship between three or more entities.

6. Etymon ID, consists of the ID of the set of etymon, donor language, and sense in the Etymon table related to this entity
7. Notes, consists of explanations or descriptions on the history of usage etc.

The entity is said to be 'minimally-designed' because (1) the number of referenced entities cannot be more than two and (2) the number of properties to be augmented to an entity are kept to be as few as possible. As explained above for A-AND and A-OR, referencing two entities is minimum to create all possible logical-etymological references. We use the term 'logical-etymological' because the reference between entities are both logical (having OR or AND) and etymological (as explained in the use cases of A-AND). Although an entity may not have any logical-etymological reference with other entities, it will always have at least one logical-etymological relationship with another entity. An entity may not have any logical-etymological reference with other entities because it is a base for other entities. Other entities will be referenced to that entity. Table 3 illustrates the relations between entities in Entity Table. In this table, the ‘Headword ID’ is replaced by ‘Headword’ for clarity purpose. ‘Etymon ID’ is replaced by ‘Etymon’, ‘Donor Language’, and ‘Sense’. ‘Etymon (original script)’ and ‘Etymon (transliteration)’ are merged in one column ‘Etymon’ to save space.

**Table 3: An example of Entity Table**

| Entity ID | Headword | Ref. Entity ID 1 | Ref. Entity ID 2 | Logic | Etymon | Donor Language | Sense |
|---|---|---|---|---|---|---|---|
| 1 | | | | | تأله (ta’allaha) | Arabic | To worship |
| 2 | ilah | 1 | | ROOT | إله<br>(ilāh) | Arabic | One who is worshipped |
| 3 | Allah | 2 | | ROOT | ﷲ<br>(Allāh) | Arabic | One who is worshipped |
| 4 | | | | | خلف<br>(khalafa) | Arabic | To replace or to represent |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | khalifah | 4 | | ROOT | خليفة<br>(khalīfah) | Arabic | Representative or leader |
| 6 | khalifatullah | 3 | 5 | AND | خليفة الله<br>(khalīfatullah) | Arabic | God’s representative |
| 7 | | | | | خار-يخير<br>(khāra) | Arabic | Exceeding, the better one |
| 8 | | | | | خار-يخور<br>(khāra) | Arabic | Mooing (animal voice) |
| 9 | | 7 | 8 | OR | استخار<br>(istakhāra) | Arabic | To request for mercy |
| 10 | | 7 | | ROOT | اختار<br>(ikhtāra) | Arabic | To choose from |
| 11 | istikharah | 9 | | ROOT | استخارة<br>(istikhārah) | Arabic | To request to be given the best among two or more choices |
| 12 | akhir | | | | آخر<br>(ākhir) | Arabic | Behind, eternal essence |

From the example above, we can make these distinctions:

1. ID 1, 4, 7, 8, and 12 are called ‘basic entities’. These entities do not have any logical-etymological reference with other entities.
2. ID 2, 3, 5, 6, 9, 10, and 11 are called ‘composite entities’. These entities have at least one logical-etymological reference.
3. ID 2, 3, 5, 6, 11, and 12 are called ‘head entities’. These entities are directly related to headwords.

In addition to database structure, there are several things that need to be considered in the verification stage, such as reliable data sources, standardized transliterations, and scripts (non-Unicode characters, variants).

## Conclusion

We have discussed the augmentation of KBBI database with etymological information. The aim is to augment the existing KBBI database with complex etymological information, especially etymological routes and logical-etymological relationships. We have designed a database structure with tables that can accommodate every possibility of etymological routes and logical-etymological relationships. However, there are some issues left, such as verification of sources by experts and workflow towards the end of the project (data input).

## Acknowledgment

We thank Ibnu Kharish for helping us with loanwords from Arabic. We gratefully acknowledge 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.

## References

Adiwimarta, Sri Sukesi, Adi Sunaryo, Saodah Nasution, Hartini Supadi, Achmad Patoni, and Umi Basiroh (1987) *Kamus Etimologi Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Harun, Ramli, Aliudin Mahyudin, and Achmad Patoni (1984) *Kamus Etimologi Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Jones, Russell (general ed.), C.D. Grijns, J.W. de Vries (eds.) (2007) *Loan-words in Indonesian and Malay*. Compiled by the Indonesian Etymological Project. Leiden: KITLV Press.

Jumariam, Meity T. Qodratillah, and C. Ruddyanto (eds.) (1996) *Senarai Kata Serapan dalam Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Kamajaya, Ian, David Moeljadi, and Dora Amalia (2017) KBBI Daring: A Revolution in The Indonesian Lexicography. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.*, Leiden, pp 513–530.

Moeljadi, David, Ian Kamajaya, and Dora Amalia (2017) Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications. In *Proceedings of The 11th International Conference of the Asian Association for Lexicography*, Guangzhou, pp 64–80.

Tadmor, Uri (2009) Loanwords in Indonesian. In Haspelmath, Martin and Tadmor, Uri (eds.) *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.