JURNAL KAJIAN BALI
Journal of Bali Studies
p-ISSN 2088-4443 # e-ISSN 2580-0698 Volume 13, Nomor 01, April 2023
Terakreditasi Sinta-2

From Corpus to Junior Dictionary: An Example of the Balinese Language

Gusti Ayu Praminatih*
Institut Pariwisata dan Bisnis Internasional
DOI: https://doi.org/10.24843/JKB.2023.v13.i01.p03

Abstract

Corpus has significantly contributed to dictionary-making. However, despite this high potential, scholars need to pay more attention to the benefits of corpus to junior dictionary development of Indonesia's local language, i.e., the Balinese language. To fill this gap, the researcher built a 56590 word-sized corpus from the data obtained from published Balinese short stories and children's songs. AntCont version 4.1.1 software was employed to retrieve words and collocations, focusing on selecting nouns for the junior dictionary entries. The study was the first of its kind to utilise corpus to design a junior dictionary for the Balinese language. Theoretically, this study significantly improved the design of a new corpus-based dictionary for junior users that entails unique and culturally bound words in Balinese. Practically, this study expands the number of dictionaries provided as a source for learning Balinese, primarily for junior users.

Keywords: corpus linguistics; junior dictionary; language preservation; lexicography

1. Introduction

When encountering an unfamiliar word or learning a new language, it is natural for the language users to consult a dictionary since it provides definitions, examples, and contexts of usage of the entry searched. However, only a few existing dictionaries provide comprehensive culturally bound terms. Consequently, it is crucial to address this issue by attempting to design a corpus dictionary. A corpus is an extensive database stored in a computer that contains the empirical descriptions and use of language sourced from written and spoken texts, i.e., daily conversations, newspapers, and even novels; thus, it could represent the language variety highly significant to the development of the study of language (Biber, 2011; Lindquist, 2009; McEnery & Hardie, 2008).

* Corresponding author: gusti.praminatih@ipb-intl.ac.id
Article submitted: 27 January 2023; Accepted: 1 April 2023

Subsequently, a corpus offers the best way to depict a textual domain, and corpus analysis is the preeminent empirical approach that is advantageous for analyzing the patterns of language use (Biber, 2011), including cultural terms.

Further, the corpus database gives enormous benefits to lexicographers by providing substantial evidence of language use and has become a potential boon for progressive new models of lexicography (Hanks, 2012; Krishnamurthy, 2006). Lexicographers strive to compile vocabularies with the compilations of databases in the form of corpora. Language dictionaries offer not only information about the language and definitions of the entries but also the correct usage and comprehension of the linguistic expression of the language (Bergenholtz & Kaufmann, 2017; Seargeant, 2011). Therefore, the benefit of corpus to dictionary making and development is indisputable.

The significance of corpus usage can be observed in several existing dictionaries. Longman Dictionary has new features such as relative frequencies of words, collocations, and grammatical patterns of spoken and written English (Kilgarriff, 1997). Earlier studies reveal that there is a significant attempt to improve the English for Specific Purposes dictionaries (Kwary, 2010, 2011b, 2011a, 2013), cultural dictionaries such as the Australian-English dictionary (Kwary & Miller, 2013) and corpus-driven dictionaries, especially in less widespread languages, i.e., the extensive studies of African languages such as Northern Zulu and Bantu (de Schryver, 2010; de Schryver & Prinsloo, 2000, 2012; de Schryver & Taljard, 2007) and India (Dash & Ramamoorthy, 2018). However, none of these studies discussed the opportunity of exploiting corpus for junior dictionaries, notably Indonesia's local languages.

Local language has always been a fascinating topic for linguists and language researchers. Preserving a local language would protect a nation's language diversity and the speakers' perceptions of their knowledge, values, and beliefs through culturally bound terms (Kwary & Miller, 2013). Ethnologue claims that Indonesia has 701 living languages (Ethnologue, 2022). Meanwhile, the Indonesian government reported that the nation's language diversity reaches about 718 local languages, including the Balinese language (Kementerian Pendidikan dan Kebudayaan Republik Indonesia, 2023). Consequently, compiling these languages into dictionaries requires significant endeavour. Moreover, the Balinese language is a dynamic, ever-changing language from its ancient form to its modern form (Beratha, 2012) and also receives many influences from other languages (Pastika, 2012), making it an excellent local language to be designed as a specific purpose or cultural dictionary, especially for users at a young age.

Recently, Indonesia's primary dictionary that entails and accommodates the existence of local languages is called *The Great Indonesian Dictionary* (KBBI

Daring, 2022), which still continuously improves. Recently, there are also several dictionaries that mainly serve the Balinese language. Recently, several dictionaries mainly serve the Balinese language. There are printed dictionaries such as *Tuttle Balinese – English Dictionary* (Shadeg, 2007) and online dictionaries such as *Kamus Bahasa Bali - Indonesia* (Balai Bahasa Provinsi Bali, 2022) and *Basa Bali Wiki* (Basa Bali Wiki, 2022). However, these existing dictionaries are designed for general users such as students, scholars and visitors. Meanwhile, the Balinese dictionary, aimed to be used for junior users, lags far behind.

The junior dictionary is essential because it has been proven to be a practical approach to helping junior language learners, significantly enhancing their reading, spelling, and phonology (Beech, 2004; de Schryver & Prinsloo, 2003). Further, the dictionary can assist vocabulary development driven mainly by the increasing number of known nouns (Segbers & Schroeder, 2017). Moreover, a previous study confirm that writing for juniors differs from adults, increasing the need to separate children's corpus (Wild et al., 2013). Hence, this study fulfils the gap of lack of effort in designing the Balinese junior dictionary.

2. Literature Review

Historically, the earliest modern dictionary-making was conducted by a British lexicographer named Samuel Johnson, who first published a dictionary called *A Dictionary of the English Language* in 1755 (Johnson, 2021). Further, another milestone in lexicography was made by Noah Webster, who compiled a dictionary called *An American Dictionary of the English Language*, published in 1828 with the addition of specific American words, i.e., *skunk* and simplified conventional spelling, i.e., *musick* to *music* in the dictionary entries (Merriam-Webster, 2023). Moreover, in its development, the dictionary, called the Merriam-Webster dictionary, has many features, including the *A Word List for Kids* (Merriam-Wesbter, 2023). Ultimately, the revolution of the dictionary that exploited corpus was conducted by John Sinclair, who created a revolutionized dictionary for learners known as *Cobuild Dictionary* in 1987 (Collins, 2023).

In the age of the Internet, there are finance dictionaries remain poor in the utilization of this technology; meanwhile, other finance dictionaries excessively use it, causing the demand to implement the modern theory of lexicography functions, focusing on the users, i.e., Indonesian finance students (Kwary, 2010, 2011b). Further, implementing the modern theory of lexicography functions is also beneficial for developing business dictionaries in smartphones that have yet to consider the users' needs (Kwary, 2013). Determining technical vocabularies for ESP dictionaries in various disciplines has also been conducted (Kwary, 2011a). Other than the dictionaries for specific purposes, there is also an attempt to investigate an online cultural dictionary database for Australian English,

presenting culturally bound terms in the form of words, phrases, sayings, signs, and symbols that can supply the variety of needs of users (Kwary & Miller, 2013).

Furthermore, several studies have been conducted regarding the dictionary of the less widespread language. For example, there is an investigation on revolutionizing the existing Zulu language dictionary using the assistance of corpus (de Schryver, 2010). Besides, there are studies regarding the macrostructure and microstructure of the dictionary of the African languages (de Schryver & Prinsloo, 2000b, 2012). Then, an attempt to compile a corpus-based dictionary of Northern Sotho is conducted by exemplifying the approach to the language mini-grammar (de Schryver & Taljard, 2007). Moreover, in the Asian context, a study reveals that the corpus of the Indian languages significantly contributes to the development of electronic dictionaries (Dash & Ramamoorthy, 2018).

The previous studies have described the journey of the earliest attempts at dictionary-making and the development of dictionaries in the internet era. The existing literature explains that the development of ESP dictionaries, cultural dictionaries, and dictionaries for a less widespread language already reach milestones. Nevertheless, the scholars seem inattentive to the junior dictionary. Junior dictionary users need to get exciting and appealing language exposure through pleasant dictionary features for them to read. Thus, this study is promising for the development of a junior dictionary.

3. Methods and Theory

3.1 Method

The corpus data for the present study were collected and analyzed in six steps. First, the researcher selected the data using short stories compilations. One of the best existing sources to support this study was a short story compilation entitled *Kumpulan Satua Bali (Dongeng Rakyat Bali)* (Suwija et al., 2019). Moreover, the more enormous the data for the corpus, the better; the researcher also added the Balinese children's songs as the data. Second, the researcher converted the data into plain text format. This step was taken because the corpus software employed in this study could only be operated using this file format. Third, after all the data were in plain text, they were created as a corpus using the AntCont version 4.1.1 software (Anthony, 2022).

Fourth, once the data were stored in the software, the researcher then retrieved the data by clicking the feature Start in the Word section (see Photo 1.) Fifth, after the data were retrieved, they must be sorted. The data sorting was required to obtain only relevant words carefully selected as junior dictionary entries. For the junior dictionary in the present study, the researcher targeted

only animal and plant-related vocabulary. Thus, vocabularies other than animal and plant were dismissed. Sixth, since this dictionary also aimed to provide additional information to the users, the collocation of each vocabulary was added. The software automatically generated these collocations.

3.2 Theories

Corpus

Scholars have studied the study of corpus linguistics for decades. This study encompasses the compilation of both spoken and written language as an extensive database to describe the nature, structure, and use of a particular language. Further, scholars define a corpus as an extensively sizeable computer-stored database containing actual descriptions and language use crucial to language studies (Biber, 2011; Lindquist, 2009; McEnery & Hardie, 2008). Consequently, the corpus offers the ultimate way to represent language most authentically; therefore, corpus analysis is a leading lucrative empirical approach to analyzing patterns of language use (Biber, 2011). Corpus databases notably benefit lexicographers by supplying substantial evidence of language use and thus beneficial for new lexicographic model proliferation (Hanks, 2012; Krishnamurthy, 2006).

Lexicography

Lexicography is a linguistic study of compiling and making a dictionary. For many years, lexicographers have endeavored to collect vocabulary using a database such as a corpus. Scholars recommend that a dictionary should offer not only information about the language and the definitions of the entries but also the correct use and understanding of the linguistic expressions of the language, which conforms to the benefit offered by corpus regarding the practical use of language (Bergenholtz & Kaufmann, 2017; Seargeant, 2011). The evidence of the significance of corpus-based frequency information in dictionaries for learners, for example, the Longman Dictionary, features emerging features such as word relative frequencies, collocations, and grammatical patterns of spoken and written English (Kilgarriff, 1997).

4. Results and Discussion

4.1 Retrieved Vocabularies

Photo 1. illustrates the preview of AntCont version 4.1.1 software and the retrieved data. In this photo, from the left part was Files, where the researcher used two files. Tokens were the size of the corpus. Below the feature Tokens were the file names of the corpus in the form of plain text format. At the top right was the Word feature, where the data were retrieved. To retrieve the

data, the researcher had to click the feature Start at the bottom right. At the left of Word was Collocate, which was used to retrieve the collocation of the word. Page Size 5000 hits indicated that the data search was around 5000 hits. After obtaining these data, the researcher needed to select only relevant vocabulary as dictionary entries meticulously; meanwhile, irrelevant data were dismissed.

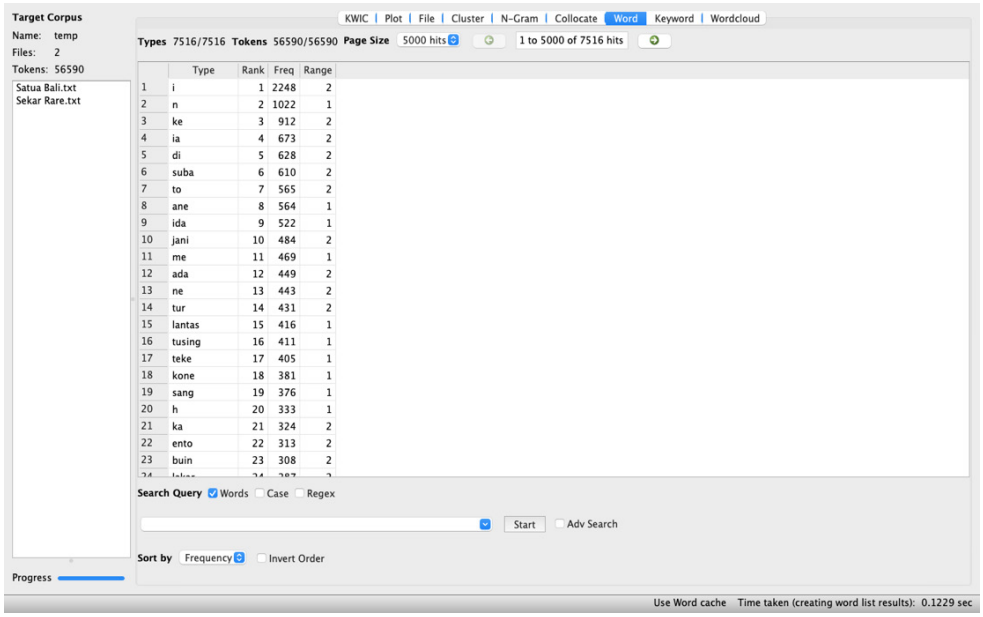

Photo 1. Data Retrieved from Antconc 4.1.1 Software (Author's Print Screen)

After selecting the relevant data, the researcher categorized the finding into two categories. First, there were 68 animal-related vocabularies, as shown in Table 1.

Table 1. Antconc Data of the Selected Animal-Related Vocabularies

No.	Data in the Balinese Language	English Translation	The Highest Collocation in the Balinese Language	English Translation
1	*lutung*	langur	*i*	a, the
2	*macan*	tiger	*poleng*	striped, black and white
3	*cicing*	dog	*gudig*	mangy
4	*kedis*	bird	*puuh*	quail
5	*balang*	grasshopper, a proper noun	*tamak*	greediness, a proper noun
6	*naga*	dragon	*Basukih*	the name of a mythical dragon

No.	Data in the Balinese Language	English Translation	The Highest Collocation in the Balinese Language	English Translation
7	*bojog*	monkey	*kedis*	bird
8	*lelipi*	snake	*gadang*	green
9	*bikul*	mouse	*semal*	squirrel
10	*kidang*	deer	*i*	a, the
11	*jaran*	horse	*nayanin*	expect
12	*kakua*	turtle	*kakua*	turtle
13	*udang*	prawn	*gede*	big
14	*cangak*	kind of heron	*cangak*	kind of heron
15	*buron*	animal	*tawah*	bizare
16	*kunang-kunang*	firefly	*kunang-kunang*	firefly
17	*lelasan*	many-stripped skink	*i*	a, the
18	*kuuk*	kind of weasel	*meng*	cat
19	*sampi*	cattle	*lelasan*	many-stripped skink
20	*alu*	water monitor lizard	*kedis*	bird
21	*puuh*	quail	*kedis*	bird
22	*singa*	lion	*i*	a, the
23	*katak*	frog	*ratun*	the king, queen
24	*crukcuk*	yellow-billed shrike	*kuning*	yellow
25	*semal*	squirrel	*bikul*	mouse
26	*sang mong*	Bali tiger	*i*	a, the
27	*garuda*	a mythical bird known as the vehicle of Vishnu	*sang*	title or respect for important or holy people
28	*kelesih*	pangolin	*nebe*	bushy backyard
29	*lembu*	ox	*i*	a, the
30	*tuma*	body louse	*titih*	bed bug
31	*beduda*	dug beetle	*jalane*	the walk
32	*titih*	bed bug	*tuma*	body louse
33	*kambing*	goat	*takutin*	afraid
34	*yuyu*	crab	*lele*	catfish
35	*blatuk*	woodpecker	*kulkul*	drum or bell made of a hollow log
36	*kancil*	mouse-deer	*i*	a, the
37	*kuluk*	puppy	*kuluk*	puppy
38	*bano*	houndfish	*be*	fish

No.	Data in the Balinese Language	English Translation	The Highest Collocation in the Balinese Language	English Translation
39	*kutu*	lice	*alihin*	please find, please search
40	*capung*	dragonfly	*bangkok*	kind of dragonfly
41	*jeleg*	kind of fish with a moustache	*be*	fish
42	*bangkung*	female pig	*tra*	not
43	*kucit*	piglet	*poleng*	striped, black and white
44	*blibis*	grouse	*blibis*	grouse
45	*sangsiah*	golden-headed cisticola	*kedis*	bird
46	*taluh*	egg	*kakul*	golden apple snail
47	*curik*	starling, myna	*curik*	starling, myna
48	*dongkang*	toad	*enjok-enjok*	limping
49	*kakul*	golden apple snail	*taluh*	egg
50	*memeri*	duckling	*madagang*	to sell, trade
51	*cekcek*	small gecko	*tomben*	rarely, infrequently
52	*kebo*	buffalo	*baka*	hypocrite
53	*legu*	mosquito	*cenik*	small
54	*semut*	ant	*semut*	ant
55	*angsa*	swan	*cai*	you (for male)
56	*asu*	dog	*blanguyunge*	a proper noun
57	*gajah*	elephant	*macanda*	play around
58	*jangkrik*	cricket	*jongkok*	squad down
59	*kupu-kupu*	butterfly	*katepukin*	seen
60	*lindung*	eel	*lele*	catfish
61	*meong*	cat	*bikul*	mouse
62	*testes*	fry	*udang*	prawn
63	*buyung*	fly	*kekembungan*	baloon
64	*clepuk*	owl	*kepet-kepet*	scratch oneself
65	*domba*	sheep	*semut*	ant
66	*goak*	crow	*kepet-kepet*	scratch oneself
67	*jair*	tilapia	*tawes*	Java barb
68	*kalejengking*	scorpion	*lelintah*	leech

Source: Processed Data from Antconc 4.1.1 Software (2022)

Second, there were 42 plant-related vocabularies, as revealed in Table 2. This study found that the plant-related vocabularies were lesser than animal-related vocabulary. The possible explanation was the nature of the data that mainly came from the Balinese children's short stories and songs, which contained more stories about animals or fable.

Table 2. Antconc Data of the Selected Plant-Related Vocabularies

No.	Data in the Balinese Language	English Translation	The Highest Collocation in the Balinese Language	English Translation
1	*punyan*	tree	*kayune*	the tree
2	*ketimun*	cucumber	*mas*	gold, a proper noun
3	*bawang*	shallot	*kesuna*	garlic
4	*kayu*	wood	*negen*	carry a load on the shoulder
5	*kesuna*	garlic	*bawang*	onion
6	*don*	leaf	*getah*	tree sap
7	*bunga*	flower	*bunga*	flower
8	*biu*	banana	*nasak*	ripe
9	*padi*	paddy	*nembuk*	pounding
10	*pudak*	pandanus flower	*luas*	to go out
11	*sekar*	flower	*ngunggahang*	to climb, to eat
12	*tiing*	bamboo	*buluhe*	reed
13	*buah*	betelnut	*lisah*	a kind of half-oil
14	*jaka*	sugar palm	*gedug*	the meaning is unidentified
15	*padang*	grass	*don*	leaf
16	*kacang*	bean	*tabia*	chili
17	*tunjung*	lotus	*beru*	blue
18	*woh-wohan*	fruit	*wangi*	fragrant
19	*bungkil*	banana stump	*tiinge*	the bamboo
20	*taru*	wood, tree, mythological tree	*gini*	name of a mythical dragon
21	*nangka*	jackfruit	*baneh*	different, foreign
22	*padma*	red lotus	*capah*	decorative part of a large fishing boat

No.	Data in the Balinese Language	English Translation	The Highest Collocation in the Balinese Language	English Translation
23	*tabia*	chilli	*tomat*	tomato
24	*tomat*	tomato	*tabia*	chilli
25	*waluh*	pumkin	*sumping*	rice cake
26	*carang*	branch, twig	*alihanga*	look for something
27	*jagung*	corn	*guungan*	cage
28	*kacang-kacangan*	beans	*tabia*	chilli
29	*madori*	giant calotrope	*getah*	tree sap
30	*ubi*	yam	*tlengis*	kind of a roasted food wrapped in a banana leaf
31	*boni*	bignay	*simarang*	the meaning is unidentified
32	*bunut*	large tree similar to the banyan tree	*simarang*	the meaning is unidentified
33	*cekuh*	galangal	*basa*	spice
34	*danyuh*	dry coconut leaf	*mirib*	resemble, apparently
35	*dapdap*	Indian coral tree	*menyan*	Sumatra benzoin tree
36	*daun*	leaf	*widuri*	giant calotrope
37	*delima*	pomegranate	*alihanga*	look for something
38	*duren*	durian	*sentok*	the meaning is unidentified
39	*gabah*	grain	*megecelan*	massage
40	*gandum*	wheat	*ngamah*	eat
41	*jepun*	frangipani	*bunga*	flower
42	*kapas*	cotton	*bebed*	bandaged

Source: Processed Data from Antconc 4.1.1 Software

4.2 Discussion

The present-day Balinese language dictionaries have been innovative. They provided word entries, phrases, idioms, expressions, derived terms with word classes and definitions, and even offered related searches. However, these dictionaries assisted students, scholars, and visitors. While this innovation was

crucial for language preservation, it neglected the dictionary's significance for junior users. Thus, the challenge was to compile dictionary entries suitable for junior users of the Balinese language. Consequently, this present study presented an essential insight into the initial compilation of entries for the junior dictionary.

This study exploited the benefits of a corpus database to obtain vocabulary, notably nouns. The corpus approach was employed given the reliable nature of the data, which was stored as empirical databases of a particular language representing the actual use of language from extensive spoken and written sources (Biber, 2011; Lindquist, 2009; McEnery & Hardie, 2008). For specific reasons, the present study built a specialised corpus of short stories and children's songs. First, it was due to the patterns of language use (Biber, 2011) offered by corpus so that the researcher would obtain the vocabulary, especially nouns, from the sources suitable for children. Second, since the Balinese short stories and children's songs contained many vocabularies about animals and plants, the expected findings of this study remained consistent with the existing junior dictionaries that also introduced animals and plants.

This corpus-based junior dictionary was unique because the Balinese language contained many culturally bound words. Culturally bound words reflected the community's values, beliefs and faith in a particular geographic area (Kwary & Miller, 2013). The findings of this study signified that the vocabularies contained several culturally bound words. First, *garuda* (a mythical bird known as the vehicle of Vishnu) was collocated with *sang* (title or respect for important or holy people). Second, *sang mong* (Bali tiger) was collocated with *i* (a, the). Third, *naga* (dragon) was collocated with *Basukih* (the name of a mythical dragon). Fourth, *macan* was collocated with *poleng* (striped, black and white).

Moreover, other culturally bound words and knowledge for junior dictionary users was that in the Balinese language, it was common to use animals and plants as human names. The study found two vocabularies that act as nouns and proper nouns. First, *balang* (grasshopper) was collocated with *tamak* (greediness, a proper noun). Second, *ketimun* (cucumber) was collocated with *mas* (gold, a proper noun). Since the data were primarily obtained from the Balinese children's short stories and songs, it was expected that the data findings would show the proper noun of the characters of the stories and also contain the original meanings of the word.

Although the study could only present nouns, their collocations included adjectives, verbs, and nouns. In this study, the example *lelipi* (snake) was collocated with *gadang* (green). *Lelipi gadang* is a common animal and a form of expression, i.e., *liep liep lipi gadang* (someone who looks kind at the surface

but has terrible or evil intentions). Subsequently, it validated that the corpus revealed the natural pattern of the language used. Thus, the findings can be potential entries for the junior dictionary by adding vocabulary, collocations, related information, and expressions (Kilgarriff, 1997). Since the junior dictionary was designed by selecting only nouns, the derivations form was not presented. However, with the additional information gained from collocation, it was possible to put cultural information such as cultural terminology, expressions, phrases, sayings, or events when designing a dictionary with cultural purposes (Kwary & Miller, 2013).

Accordingly, the present study confirmed the previous studies that explain the potential of employing corpus could significantly benefit the lexicographer in compiling a progressive dictionary (Hanks, 2012; Krishnamurthy, 2006). In this study, the progressive part gives additional information to enrich junior dictionary users' cultural knowledge and awareness. The findings also explained the significance of collocation in the dictionary (Sandro, 2008). Thus, in the junior dictionary, the collocations help understand the word when consulting the dictionary.

Further, the findings aligned with scholars who defined that a dictionary should give definitions and comprehensive aspects of the language, including background and foreground knowledge (Bergenholtz & Kaufmann, 2017; Seargeant, 2011). The junior dictionary designed in this study already fulfilled the requirement to include background and foreground knowledge. Then, this study presented the example of the Balinese junior dictionary entry illustrated in Photo 2.

Picture	
Entry	naga
Definition	dragon
Collocation	Basukih
Cultural information	Basukih is a mystical dragon believed to keep the balance of nature.

Photo 2. The Example of Balinese – English Junior Dictionary Entry

To the best of the researcher's knowledge, the present study was the first of its kind to develop a dictionary aimed at junior users, which was accomplished by utilising corpus and thus could give added value to the information of the dictionary. Consequently, this study filled the gap and could be complementary to the corpus-driven approach in less widespread languages such as Northern Zulu and Bantu (de Schryver, 2010; de Schryver & Prinsloo, 2000, 2012; de Schryver & Taljard, 2007) and India (Dash & Ramamoorthy, 2018). Further, it also contributed to the effort of preserving the Balinese language together with existing dictionaries such as *The Great Indonesian Dictionary* (KBBI Daring, 2022), *Tuttle Balinese – English Dictionary* (Shadeg, 2007), *Kamus Bahasa Bali - Indonesia* (Balai Bahasa Provinsi Bali, 2022), and *Basa Bali Wiki* (Basa Bali Wiki, 2022).

Ultimately, this study could trigger junior users and parents to use the Balinese language from a very early age as scholars revealed that a dictionary is a practical approach to assist junior language learners in studying a language (Beech, 2004; de Schryver & Prinsloo, 2003). Moreover, the current study attempted to compile nouns which helped increase vocabulary development (Segbers & Schroeder, 2017). Moreover, to the best of the author's knowledge, this is the first attempt to compile vocabularies, especially nouns using a corpus-driven approach for junior dictionary, following the recommendation of scholars that children and adults should be using separate corpus for their different needs (Wild et al., 2013).

5. Conclusion

A corpus offers an excellent database for the lexicographer to decide vocabularies, frequencies, collocations, and other related information that can be selected for a dictionary. The present study leads to selecting vocabularies as junior dictionary entries by utilizing corpus. It is an example of the development of the dictionary for junior users by providing easy-to-understand entries, which are completed with collocations and cultural information. Accordingly, future researchers and lexicographers are expected to develop junior dictionaries for children in other cultures by exploiting valuable resources like corpus. Although the present study already revealed the potential of compiling vocabularies, frequencies, and collocations that also entailed culturally bound words and expressions, future research is required to formulate definitions and compile other parts of speech, such as verbs, adjectives, and adverbs. Consequently, a larger corpus is urgently needed to complete this junior dictionary.

Acknowledgement

An earlier version of this article was presented at The International Conference on Local Languages (ICLL) 2023. The researcher would like to

express her gratitude to Asosiasi Peneliti Bahasa-Bahasa Lokal (APBL) as the organizing committee of ICLL 2023 and *Jurnal Kajian Bali*, which had been the sponsor of the conference. The conference was held in a hybrid from the Universitas Warmadewa, Denpasar, on 17 February 2023.

Bibliography

Anthony, L. (2022). *Laurence Anthony's AntConc*. https://www.laurenceanthony.net/software/antconc/

Balai Bahasa Provinsi Bali. (2022). *Kamus Bahasa Bali - Indonesia*. http://kamusbahasaprovinsibali.id/bali-indonesia/cari

Basa Bali Wiki. (2022). *Virtual Dictionary of Balinese Language, Culture, Tradition, History, and Arts - BASAbali Wiki*. https://dictionary.basabali.org/Main_Page

Beech, J. R. (2004). Using a dictionary: Its influence on children's reading, spelling, and phonology. *Reading Psychology*, *25*(1), 19–36. https://doi.org/10.1080/02702710490271819

Beratha, N. L. S. (2012). Frasa Bahasa Bali Kuna dan Perkembangannya ke Bahasa Bali Modern. *Jurnal Kajian Budaya*, *02*(02), 69–86.

Bergenholtz, H., & Kaufmann, U. (2017). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. *HERMES - Journal of Language and Communication in Business*, *10*(18), 91. https://doi.org/10.7146/hjlcb.v10i18.25413

Biber, D. (2011). Corpus linguistics and the study of literature. *Scientific Study of Literature*, *1*(1), 15–23. https://doi.org/10.1075/ssol.1.1.02bib

Collins. (2023). *The History of Cobuild - Collins Dictionary Language Blog*. Collins. https://blog.collinsdictionary.com/the-history-of-cobuild/

Dash, N. S., & Ramamoorthy, L. (2018). *Utility and Application of Language Corpora*. Springer. https://doi.org/10.1007/978-981-13-1801-6

de Schryver, G. M. (2010). Revolutionizing Bantu lexicography - A Zulu case study. *Lexikos*, *20*, 161–201. https://doi.org/10.4314/lex.v20i1.62689

de Schryver, G. M., & Prinsloo, D. J. (2000a). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages*, *20*(4), 310–330. https://doi.org/10.1080/02572117.2000.10587438

de Schryver, G. M., & Prinsloo, D. J. (2000b). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages*, *20*(4), 310–330. https://doi.org/10.1080/02572117.2000.10587438

de Schryver, G. M., & Prinsloo, D. J. (2012). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. *Http://Dx.Doi.Org/10.1080/02572117.2000.10587437*, *20*(4), 291–309. https://doi.org/10.1080/02572117.2000.10587437

de Schryver, G. M., & Taljard, E. (2007). Compiling a Corpus-based dictionary grammar: An example for Northern Sotho. *Lexikos*, *17*, 37–55. https://doi.org/10.5788/17-0-1163

de Schryver, G.-M., & Prinsloo, D. J. (2003). Compiling a lemma-sign list for a specific target user group: The Junior Dictionary as a case in point. *Dictionaries: Journal of the Dictionary Society of North America*, *24*(1), 28–58. https://doi.org/10.1353/dic.2003.0014

Ethnologue. (2022). *Languages of Indonesia | Ethnologue*. SIL International. https://www.ethnologue.com/product/19-Report-ID

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398–436. https://doi.org/10.1093/ijl/ecs026

Johnson, S. (2021). Preface to A Dictionary of the English Language (1755). In *Yale University*. Yale University Press. https://doi.org/10.12987/9780300258004-038/HTML

KBBI Daring. (2022). *Halaman Statistik*. Kementerian Pendidikan, Kebudayaan, Riset, Dan Teknologi. https://kbbi.kemdikbud.go.id/Beranda/Statistik

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2023). *Bahasa dan Peta Bahasa di Indonesia*. https://petabahasa.kemdikbud.go.id/

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*(2), 136–155. https://doi.org/10.1093/ijl/10.2.135

Krishnamurthy, R. (2006). Corpus Lexicography. *Encyclopedia of Language & Linguistics*, *1999*, 250–254. https://doi.org/10.1016/B0-08-044854-2/00416-8

Kwary, D. A. (2010). Access routes of internet finance dictionaries: Present solutions and future opportunities. *Lexikos*, *20*, 272–289. https://doi.org/10.4314/lex.v20i1.62715

Kwary, D. A. (2011a). A hybrid method for determining technical vocabulary. *System*, *39*(2), 175–185. https://doi.org/10.1016/j.system.2011.04.003

Kwary, D. A. (2011b). Adaptive hypermedia and user-oriented data for online dictionaries: A case study on an English dictionary of finance for Indonesian students. *International Journal of Lexicography*, *25*(1), 30–49. https://doi.org/10.1093/ijl/ecr008

Kwary, D. A. (2013). Principles for the Design of Business Dictionaries on Mobile Applications. *Hermes – Journal of Language and Communication in Business*, *50*, 69–81.

Kwary, D. A., & Miller, J. (2013). A model for an online Australian English cultural dictionary database. *Terminology*, *19*(2), 258–276. https://doi.org/10.1075/term.19.2.05kwa

Lindquist, H. (2009). Corpus linguistics and the description of English. *Corpus Linguistics and the Description of English*, 1–219. https://doi.org/10.2478/ICAME-2020-0006

McEnery, T. H., & Hardie, A. (2008). *Corpora in Linguistics*. 1–312.

Merriam-Webster. (2023). *About Us | Merriam-Webster*. Merriam-Webster. https://www.merriam-webster.com/about-us/americas-first-dictionary

Merriam-Wesbter. (2023). *A Words List for Kids: Browse the Student Dictionary | Merriam-Webster*. Merriam-Webster. https://www.merriam-webster.com/browse/kids/

Pastika, I. (2012). Pengaruh Bahasa Asing terhadap Bahasa Indonesia dan Bahasa Daerah: Peluang atau Ancaman? *Jurnal Kajian Bali (Journal of Bali Studies)*, *2*(2), 141–164.

Sandro, N. (2008). The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos*, *18*(0), 170–189. http://lexikos.journals.ac.za/pub/article/view/483/179

Seargeant, P. (2011). Lexicography as a philosophy of language. *Language Sciences*, *33*(1), 1–10. https://doi.org/10.1016/j.langsci.2010.06.002

Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297–320. https://doi.org/10.1177/0265532216641152

Shadeg, N. A. (2007). *Tuttle Balinese-English Dictionary*. Tuttle Publishing. https://books.google.com/books/about/Tuttle_Balinese_English_Dictionary.html?hl=id&id=20ktBAAAQBAJ

Suwija, I. N., Darmada, I. M., & Mulyawan, I. N. R. M. (2019). *Kumpulan Satua (Dongeng Rakyat Bali)* (1st ed.). Pelawa Sari.

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography*, *26*(2), 190–218. https://doi.org/10.1093/ijl/ecs017

Author's Profile

Gusti Ayu Praminatih is a linguist and lecturer at Institut Pariwisata dan Bisnis Internasional, Denpasar, Bali, Indonesia. She earned her degree in Linguistics from Universitas Airlangga in 2018. Her research interest includes language and gender, corpus linguistics, and lexicography.