The International Journal of Social Sciences World
TIJOSSW is Available Online at:
https://www.growingscholar.org/journal/index.php/TIJOSSW
Vol. 6 No. 1, January – June 2024, pages: 225- 237
DOI: https://doi.org/10.5281/zenodo.12621570
ISSN **2690-5167**
Growingscholar Publisher

Selecting Representative Data Samples from Corpus Collection with Specific Target Domains

Ni Made Ayu Widiastuti [1], **Ketut Artawa** [2], **I Made Rajeg** [3],
I Nyoman Udayana [4], **Gede Primahadi Wijaya Rajeg** [5]

Article history:
Received: May 18, 2024; Accepted: June 30. 2024; Displayed Online: June 30, 2024; Published: June 30, 2024

Keywords

Representative sample;
Target domain;
Availability;
Procedures;
Corpus;

Abstract

This study aims to show the availability of corpus files containing tokens generated from the search of the keywords and to select representative data samples based on the availability of the tokens. This paper reported the case study of corpus linguistics data collection procedures as careful considerations. The *Leipzig Corpora Collection* (LCC) in Indonesian is the data source. The corpus sampling frame was adapted by considering the structured guidelines used to define the criteria for selecting balanced and representative samples. There are two main procedures to get the representative data samples, i) to know the availability of corpus files containing tokens generated from the search of the keywords; and ii) to select the representative data samples based on the availability of the tokens. The results are ten files in ten years between 2013–2022, each of them has ±1,000 linguistic expressions as the balanced and representative data samples.

1. Introduction

A corpus is a large organized set of texts used in linguistic research that is typically stored and processed electronically these days. It may include written texts, spoken language transcriptions, or both. A corpus contains raw data that have not been annotated or linguistically marked, so the

[1] Udayana University, Denpasar, Indonesia. *Email:* ayu_widiastuti@unud.ac.id
[2] Udayana University, Denpasar, Indonesia. *Email:* ketut_artawa@unud.ac.id
[3] Udayana University, Denpasar, Indonesia. *Email:* made_rajeg@unud.ac.id
[4] Udayana University, Denpasar, Indonesia. *Email:* nyoman_udayana@unud.ac.id
[5] Udayana University, Denpasar, Indonesia. *Email:* primahadi_wijaya@unud.ac.id

researcher will manually annotate all the necessary data (Gries, 2009:9). An empirical study of language can be conducted using a corpus.

The analysis of language used in communication in large quantities using a computer is called corpus linguistics (McEnery and Hardie, 2012). The field of corpus linguistics deals with the linguistic analysis of electronic text collections (Lusta et al., 2023). In terms of a particular language, it studies a language-dependent field (Stoykova, 2014). Researchers can examine patterns, frequencies, and variances in vocabulary and linguistic structures by investigating language use. According to Mizin & Slavova (2023), corpus study enables researchers to obtain objective results.

In corpus linguistics, the use of language can be specific to its types. Researchers can choose certain types of corpora related to their research. General corpora that contain a diverse range of text types are the Corpus of Contemporary American English (COCA)[6], the British National Corpus (BNC)[7], and the Leipzig Corpora Collection (LCC)[8]. The examples of specialized corpora that concentrate on particular fields, registers, or genres are ComPara which is a dataset in English containing computational architecture (see Horvath, 2022), Balance Corpus of Contemporary Written Japanese[9] (see Oeinada et al., 2021), and legal corpora (see Gries & Gales, 2024). In terms of the spoken corpora that contain a transcription of spoken language, the Santa Barbara Corpus of Spoken American English (SBCSAE)[10], and historical corpora that include texts from previous eras such as the Helsinki Corpus of English Texts (HCET)[11] are several examples.

Since corpus is the source of big data, selecting representative data samples is challenging for researchers. The representativeness and balance are matters of judgment and are therefore only approximate (Kennedy, 1998), and their measurements are matters of degree (McEnery & Hardie, 2012, p. 10). It takes effort to make sure that the samples adequately represent the entire diversity and variability of the dataset when dealing with modern corpora that comprise millions or even billions of tokens distributed across many files. The sampling procedure is further complicated by the fact that language changes over time, requiring that researchers consider these changes.

To achieve a balanced sample, careful considerations and exact statistical techniques are required to deal with such massive and dynamic data. Egbert et al. (2020) state that there is a set of procedures in the process of collecting samples of data from a corpus. It may include choosing and assessing corpora, formulating linguistically motivated research questions, choosing observational units and variables, choosing variables that can be understood by language, comprehending and assessing the tools for corpus software that are currently available, and using minimally sufficient statistical methods.

There is a positive trend toward growth and development in corpus linguistics, despite the fact that there are still relatively few corpus studies written by Indonesian researchers, either those who only applied the qualitative method, or combined the quantitative and qualitative methods (see Yuditha, 2013; Lapasau et al., 2019; Huszka, 2020; Endarto, 2020; Oktavianti & Sarage, 2021). Moreover, a limited number of studies were carried out that comprehensively describe how to obtain representative data samples (see Rajeg, 2013; Rajeg, 2019 for examples).

This research shares the experience of selecting representative data samples from a corpus with careful consideration. It tries to capture a wide range of language use based on the availability of tokens in each file provided in the existing corpus. The specific aims of this research are to show the

[6] COCA https://www.english-corpora.org/coca/
[7] BNC http://www.natcorp.ox.ac.uk/
[8] LCC https://wortschatz.uni-leipzig.de/en
[9] BCCWJ https://chunagon.ninjal.ac.jp
[10] SBCSAE https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[11] HCET https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/

availability of corpus files containing tokens generated from the search of the keywords and to select representative data samples based on the availability of the tokens.

2. Materials and Methods

The *Leipzig Corpora Collection* (LCC) is the source where the representative data samples are obtained. LCC is a freely accessible resource for corpora and corpus statistics (Biemann et al., 2007, p. 1). This corpus is the existing one, and recently, it includes 250 languages. Indonesian is one of the languages available, and can be downloaded from https://wortschatz.uni-leipzig.de/en/download/Indonesian. To generate the concordances from the files of the corpus, the AntConc program (Anthony, 2022) was used.

McEnery & Hardie (2012, p. 8) suggest that the corpus sampling frame is used to seek balanced and representative samples in a corpus. They state "The population is the notional space within which language is being sampled". Therefore, for example, if researchers want to figure out the interaction of language of service in stores in a specific country and a range of specific years, the sampling frame is only the corpus data that represents service interactions in a specific country and a certain range of years. However, determining the sampling frame from the corpus is not enough, it is necessary to gather relatively context-specific lexis that seems to occur more often.

In this research, the corpus sampling frame was adapted by considering the structured guidelines used to define the criteria for selecting representative samples. First, the files displayed in the LCC in Indonesian must be known. The information about the genre of texts provided, the number of files, and the range of years must be noticed. Second, the latest ten-year files are chosen and downloaded. The downloaded files are in *.rar* because each of them has 7 different .txt files and 1 SQL file. Figure 1 shows an example of the *unrar* files contained in one corpus file.

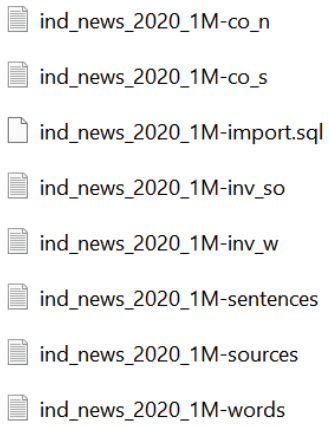


Figure 1. An example of downloaded *unrar* files

There are five pieces of information on each of the *unrar* files, they are i) '*ind*', the Indonesian language; ii) '*news*', the type of text, in this example, the texts are taken from the news; iii) 2020, the year when the texts are published; iv) 1M, the number of words in the texts; v) category of parsing, such as sentences, words, etc. Third, the files are added one by one into the AntConc to show the availability of representative data samples required for the research. The category of parsing chosen

is the 'sentences' because context is needed. Fourth, the keywords are typed in the search box in AntConc.

In this case, because the further analysis is about the conceptual metaphors of *ketakutan* 'fear' and its synonyms, of course, the synonyms have been determined based on a literature review (Shaver et al., 2001; Mulyadi, 2001; Yosephine, 2022). The synonyms of the noun *ketakutan* 'fear' in Indonesian are *kekawatiran/kekhawatiran* 'worry' (*kekuatiran* and *kekwatiran* as the variations are also included)*, kewaswasan* 'worry' (informal form based on Indonesian dictionary)*, kerisauan* 'anxiety', *kecemasan* 'anxiety'*, kegalauan* 'anxiety', *keresahan* 'anxiety', *kengerian* 'anxiety'*, ketercekaman* 'fear' (informal form based on Indonesian dictionary)*, keseraman* 'scary', *kegelisahan* 'anxiety'*, kebingungan* 'confusion', *kegugupan* 'nervousness', *kegrogian* 'nervousness', dan *kepanikan* 'panic'. Those 18 emotion words are called the target domain as the keywords. When searching the keywords using the AntConc program, several settings are necessary to gather the relatively context-specific lexis that collects words that are commonly used within a particular context or domain. There are four settings, those are *Regex:* \b(?i)noun\b, *Sort options:* 1L, 1R, 2R, *Results set: All hits* or *Random hits,* and *Context size:* 20 *tokens*. In this case, the software shows the search results in the *KeyWord in Context* (KWIC) format, and generates the concordance lines. The last step was to get representative data samples of around 1.000 tokens generated from one of the settings in AntConc, that is *All hits* or *Random hits*. This number is the consideration related to similar previous research that has been conducted (see Stefanowitsch, 2006 and Rajeg, 2019).

3. Results and Discussions

The selection of the representative data samples or linguistic expressions in the LCC Indonesian was comprised of two main procedures. Those are the results of the careful considerations based on the main topic of study, that is the distinctive metaphors of *ketakukan 'fear'* and its 6 synonyms. Before the analysis of distinctive metaphors, the data collection was done carefully and reported in this paper. The first procedure is to know the availability of corpus files containing tokens generated from the search of the keywords. In this case, the LCC Indonesian displays the files containing the sentences in Indonesian taken from various sources, such as websites, news, or texts. The sentences are called linguistic expressions. The linguistic expressions are the raw data. To get the expressions that can be called tokens with the keywords *ketakukan 'fear'* and its 6 synonyms, it needs the AntConc software to search them, however, the results are not called the balanced and representative data samples yet. Then, the second procedure is to select the representative data samples based on the availability of the tokens. This step is necessary to be done because each file has a different range of total number of tokens. The second filter was done to have a nearly balanced and representative number of tokens.

The Availability of Corpus Files Containing Tokens Generated from Search of The Keywords

In early 2023, LCC in the Indonesian language had 41 files categorized based on 20 types of text sources. The files are also differentiated based on the year and number of sentences. All the files displayed on the website (https://wortschatz.uni-leipzig.de/en/download/Indonesian), and retrieved in April 2023 are rewritten in Table 1.

Tabel 1
List of Corpus Files in LCC in 2023

No	Text's Sources_Year	Country	Number of Sentences in Each File
1	Mixed_2012	Indonesia	10K 30K 100K 300K 1M
2	Mixed_2013	Indonesia	10K 30K 100K 300K 1M
3	Mixed-tufs4_2012	Indonesia	10K 30K 100K 300K 1M
4	News_2008	Indonesia	10K 30K 100K 300K
5	News_2009	Indonesia	10K 30K 100K 300K
6	News_2010	Indonesia	10K 30K 100K 300K
7	News_2011	Indonesia	10K 30K 100K 300K
8	News_2012	Indonesia	10K 30K 100K 300K
9	News_2019	Indonesia	10K 30K 100K 300K 1M
10	News_2020	Indonesia	10K 30K 100K 300K 1M
11	News_2022	Indonesia	10K 30K 100K 300K 1M
12	News-tufs10_2011	Indonesia	10K 30K 100K 300K
13	News-tufs11_2012	Indonesia	10K 30K 100K 300K
14	News-tufs7_2008	Indonesia	10K 30K 100K 300K
15	News-tufs8_2009	Indonesia	10K 30K 100K 300K
16	News-tufs9_2010	Indonesia	10K 30K 100K 300K
17	Newscrawl_2011	Indonesia	10K 30K 100K 300K 1M
18	Newscrawl_2012	Indonesia	10K 30K 100K 300K 1M
19	Newscrawl_2015	Indonesia	10K 30K 100K 300K
20	Newscrawl_2016	Indonesia	10K 30K 100K 300K 1M
21	Newscrawl-tufs5_2011	Indonesia	10K 30K 100K 300K 1M 3M
22	Newscrawl-tufs6_2012	Indonesia	10K 30K 100K 300K 1M 3M
23	Web_2011	Indonesia	10K 30K 100K 300K
24	Web_2012	Indonesia	10K 30K 100K 300K 1M
25	Web_2013	Indonesia	10K 30K 100K 300K 1M
26	Web_2015	Brunei	10K
27	Web_2015	India	10K 30K 100K 300K 1M
28	Web_2015	Indonesia	10K 30K 100K 300K 1M
29	Web_2017	Indonesia	10K 30K 100K 300K 1M
30	Web_2018	com	10K 30K 100K 300K 1M
31	Web-public_2017	Indonesia	10K 30K 100K 300K 1M
32	Web-tufs12_2011	Indonesia	10K 30K 100K 300K
33	Web-tufs13_2012	Indonesia	10K 30K 100K 300K 1M 3M
34	Web-tufs2_2013	Indonesia	10K 30K 100K 300K 1M
35	Web-tufs3_2015	Indonesia	10K 30K 100K 300K 1M 3M
36	Wikipedia_2010	Indonesia	10K 30K 100K 300K
37	Wikipedia_2014	Indonesia	10K 30K 100K 300K 1M
38	Wikipedia_2016	Indonesia	10K 30K 100K 300K 1M
39	Wikipedia_2021	Indonesia	10K 30K 100K 300K 1M
40	Wikipedia-tufs14_2016	Indonesia	10K 100K 300K 1M
41	Wikipedia-tufs16_2016	Indonesia	30K

From Table 1, it can be seen that the display of the raw number of data in each file ranges from 10 thousand (10K) to 3 million (3M) sentences. In this case, not all of the files were downloaded. There are two considerations in choosing the files, those are: i) the highest number of raw sentences were chosen to get more variations of linguistic expressions, and ii) the country label which is not Indonesian was not included. The results of the filter based on those two considerations are shown in Table 2.

Table 2
List of Files Filtered Based on the Highest Number of Sentences in Each File and the Indonesian Sources

No	Text's Sources_Year
1	ind_mixed_2012_1M
2	ind_mixed_2013_1M
3	ind_mixed-tufs4_2012_1M
4	ind_news_2019_1M
5	ind_news_2020_1M
6	ind_news_2022_1M
7	ind_newscrawl_2011_1M
8	ind_newscrawl_2012_1M
9	ind_newscrawl_2016_1M
10	ind_newscrawl-tufs5_2011_3M
11	ind_newscrawl-tufs6_2012_3M
12	ind_web_2012_1M
13	ind_web-tufs2_2013_1M
14	ind_web-tufs3_2015_3M
15	ind_web-tufs13_2012_3M
16	ind_wikipedia_2014_1M
17	ind_wikipedia_2016_1M
18	ind_wikipedia_2021_1M
19	ind_wikipedia-tufs14_2016_1M
20	ind-com_web_2018_1M
21	ind-id_web_2013_1M
22	ind-id_web_2015_1M
23	ind-id_web_2017_1M
24	ind-id_web-public_2017_1M

The files' names were automatically generated as seen in Table 2. Each file has the 4 information separated by underscores. For example, 'ind_mixed_2012_1M' means: i) the data is in Indonesian language ('ind'), ii) types of sources are mixed (used text materials that were taken from a mixture of sources like news material, Web text, etc.), iii) the year of the sources' compilation, and iv) the number of sentences in each file, i.e. one or three million.

To get the number of tokens based on the keywords in each file, each emotion word as the target domain was searched. This step was done in AntConc with four settings: *Regex:* \b(?i)noun\b, *Sort options:* 1L, 1R, 2R, *Results set: All hits* or *Random hits,* and *Context size:* 20 *tokens*. The display of AntConc can be seen in Figure 2, and the results are shown in Table 3a and Table 3b.

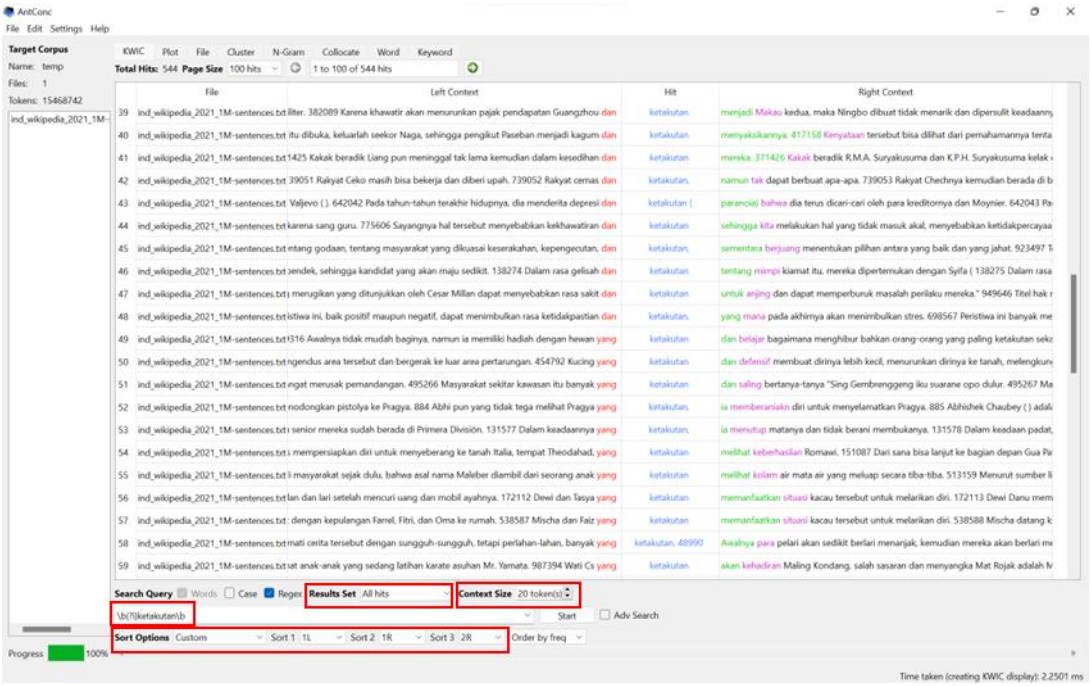


Figure 2. Display Sample of a Target Domain Searching in the AntConc

Table 3a
Number of Tokens Found in Each File with the Target Domains (a–i)

No.	File Name	Total hits/ token								
		a	b	c	d	e	f	g	h	i
		ketakutan	*kekhawatiran*	*kebingunan*	*kecemasan*	*keresahan*	*kepanikan*	*kegelisahan*	*kegalauan*	*kekuatiran*
1	ind_mixed_2012_1M	1,373	483	409	255	110	57	183	31	68
2	ind_mixed_2013_1M	625	569	226	233	108	73	125	49	28
3	ind_mixed-tufs4_2012_1M	1,003	421	398	319	89	86	203	28	51
4	ind_news_2019_1M	443	463	180	127	121	62	68	24	10
5	ind_news_2020_1M	564	860	228	236	166	188	53	19	3
6	ind_news_2022_1M	408	633	192	141	129	125	56	17	5
7	ind_newscrawl_2011_1M	624	644	265	133	159	148	75	17	23
8	ind_newscrawl_2012_1M	557	1,154	204	168	185	164	89	28	29
9	ind_newscrawl_2016_1M	403	556	159	141	163	79	110	38	27

10	ind_newscrawl-tufs5_2011_3M	1,924	1,893	784	420	555	508	234	51	67
11	ind_newscrawl-tufs6_2012_3M	1,614	3,379	630	485	526	463	1	110	77
12	ind_web_2012_1M	940	397	324	277	106	78	176	46	63
13	ind_web-tufs2_2013_1M	421	478	234	195	54	73	94	47	36
14	ind_web-tufs3_2015_3M	1,251	1,368	224	692	98	55	97	115	34
15	ind_web-tufs13_2012_3M	2,994	1,237	370	867	116	93	232	110	168
16	ind_wikipedia_2014_1M	587	215	253	96	25	74	47	16	9
17	ind_wikipedia_2016_1M	559	219	258	98	26	72	41	14	10
18	ind_wikipedia_2021_1M	554	291	266	140	46	107	47	12	5
19	ind_wikipedia-tufs14_2016_1M	595	210	254	93	36	70	46	18	10
20	ind-com_web_2018_1M	533	315	343	308	59	54	96	36	19
21	ind-id_web_2013_1M	451	491	228	194	57	73	96	46	38
22	ind-id_web_2015_1M	453	476	216	222	82	61	90	36	17
23	ind-id_web_2017_1M	426	366	258	272	91	46	87	42	17
24	ind-id_web-public_2017_1M	395	366	219	242	91	50	98	42	16
	TOTAL	**19,697**	**17,484**	**7,122**	**6,354**	**3,198**	**2,859**	**2,444**	**992**	**830**

Table 3b
Number of Tokens Found in Each File with the Target Domains (j–r)

		Total hits/ token								
		j	k	l	m	n	o	p	q	r
No.	**File Name**	*kengerian*	*kerisauan*	*kekawatiran*	*kegugupan*	*keseraman*	*kekwatiran*	*kewaswasan*	*ketercekaman*	*kegrogian*
1	ind_mixed_2012_1M	69	19	7	11	2	-	0	0	0
2	ind_mixed_2013_1M	17	23	13	4	1	-	0	0	0
3	ind_mixed-tufs4_2012_1M	35	17	9	7	3	1	0	0	1
4	ind_news_2019_1M	13	8	1	3	1	1	0	0	0
5	ind_news_2020_1M	9	10	1	3	0	1	0	0	0
6	ind_news_2022_1M	23	2	-	6	3	-	0	0	0
7	ind_newscrawl_2011_1M	8	15	2	6	1	-	0	0	0
8	ind_newscrawl_2012_1M	16	16	7	3	1	-	0	1	0
9	ind_newscrawl_2016_1M	9	8	6	1	0	1	0	0	0

10	ind_newscrawl-tufs5_2011_3M	30	55	12	19	3	2	0	1	0
11	ind_newscrawl-tufs6_2012_3M	43	52	24	14	0	1	0	1	0
12	ind_web_2012_1M	39	32	1	9	0	-	0	0	0
13	ind_web-tufs2_2013_1M	16	8	14	3	1	1	0	0	0
14	ind_web-tufs3_2015_3M	15	31	16	4	1	2	0	1	1
15	ind_web-tufs13_2012_3M	36	56	37	9	5	11	0	0	1
16	ind_wikipedia_2014_1M	27	3	-	2	2	-	1	0	0
17	ind_wikipedia_2016_1M	32	5	-	1	2	-	1	0	0
18	ind_wikipedia_2021_1M	23	4	3	3	2	-	1	0	0
19	ind_wikipedia-tufs14_2016_1M	34	4	1	2	2	-	1	0	0
20	ind-com_web_2018_1M	23	10	-	10	2	-	0	0	0
21	ind-id_web_2013_1M	18	8	11	2	1	1	0	0	0
22	ind-id_web_2015_1M	17	6	5	5	0	1	0	0	0
23	ind-id_web_2017_1M	22	11	3	5	0	-	0	0	1
24	ind-id_web-public_2017_1M	16	12	5	4	3	1	0	0	0
	TOTAL	**590**	**415**	**178**	**136**	**36**	**24**	**4**	**4**	**4**

Tables 3a and 3b show the number of tokens that have been found through the searching of each target domain (a–r) using the AntConc software and were sorted from the highest to lowest total number of tokens. The results reveal vastly different numbers of tokens.

Selecting The Representative Data Samples Based on The Availability of The Tokens

In this research, to gather the balance and representative linguistic expressions, it is necessary to get an equal number of data from each file. This technique was conducted based on the results of the number of tokens/ linguistic expressions found in each file with the target domains in Tables 3a and 3b. The files containing the target domains (a–g) are more than 2,000, therefore, they meet the criteria as the sources of linguistic expressions for this research because they occur more often than the other target domains, and there was another step to choose the balance tokens. Those target domains are *ketakutan* 'fear', *kekhawatiran* 'worry', *kebingungan* 'confusion', *kecemasan* 'anxiety', *keresahan* 'anxiety', *kepanikan* 'panic', and *kegalauan* 'anxiety'.

To get the balance tokens, there are two considerations: i) choose one of the files within the same year containing a nearly similar number of tokens of the target domain found, and ii) the files must be published in the last ten-year period, in this case, the range of year chosen was from 2013 – 2022 because the files available on the website in 2023 when the data collection was started were up to 2022. Related to the first criterion, there are three files compiled in 2013 and 2016, and two files compiled in 2015 and 2017. One of the files that does not contain a vastly different number of tokens of the target domain was chosen. Each file compiled in 2014, 2017, 2018, 2019, 2020, 2021, and 2022 did not occur more than once, therefore, all of them were chosen. In terms of the second criterion, there are 8 files eliminated because they were compiled in 2011 and 2012. The results of selecting the balance and representative files in the ten years are shown in Table 4.

Table 4
List of Files in Ten years

No.	File Name	Total hits/ token						
		a	b	c	d	e	f	g
		ketakutan	*kekhawatiran*	*kebingunan*	*kecemasan*	*keresahan*	*kepanikan*	*kegelisahan*
1	ind_web-tufs2_2013_1M	421	478	195	234	54	73	94
2	ind_wikipedia_2014_1M	587	215	96	253	25	74	47
3	ind_web-tufs3_2015_3M	1,251	1,368	692	224	277	166	262
4	ind_newscrawl_2016_1M	403	556	141	159	163	79	110
5	ind-id_web_2017_1M	426	366	272	258	91	46	87
6	ind-com_web_2018_1M	533	315	308	343	59	54	96
7	ind_news_2019_1M	443	463	127	180	121	62	68
8	ind_news_2020_1M	564	860	236	228	166	188	53
9	ind_wikipedia_2021_1M	554	291	140	266	46	106	47
10	ind_news_2022_1M	408	633	141	192	129	125	56
	TOTAL	**5,590**	**5,545**	**2,348**	**2,337**	**1,131**	**973**	**920**

The total number of tokens in each file after the selection of the ten-year period still shows a great difference, therefore, a second selection is needed. In this case, the random sampling was not taken manually, but using the AntCont with the setting called *Results set: Random hits.* For the target domain *ketakutan, kekhawatiran,* and *kecemasan,* the *Random hits* sampling was set to get 100 tokens in each file, so that 100 tokens x 10 files equals 1.000 tokens as the balanced representative samples. However, there are some numbers of tokens less than 100 in each file, then the setting *Results set: All hits* was applied to get all the tokens, but the *Results set: Random hits* was still applied for more than 100 tokens. It can be said that for the target domain *kebingungan*, *kepanikan,* and *kegelisahan,* both the *Results set: Random hits* and *All hits* were applied. For *keresahan, All hits* setting was applied, but because the tokens are more than 1.000, that is 1.131 in total, the number of tokens compiled in 2015, 2016, 2019, 2020, and 2022 were once again subtracted for 15%, and 16.3% (only the file compiled in 2015) by using random sampling technique in Excel to get the maximum number of 1.000 tokens. The results of the representative linguistic expression sample from the 10 years of files are now ±1.000 tokens. It cannot reach the same number of tokens, that is 1.000 for each file because the techniques cannot be applied in the same way for each file due to the variation range of tokens available. The results of the final sampling technique are shown in Table 5.

Table 5
List of Files in the Ten Years and ±1,000 'Random/ All Hits' Selections

No	File Name	Sampling Results (Total hits) Generated from AntConc													
		ketakutan		*kekhawatir -an*		*kecemasan*		*kebingunan*		*keresahan*		*kepanik -an*		*kegelisah -an*	
		All hits	Ran-dom hits	All hits	Ran-dom hits	All hits	Ran-dom hits & All hits (2014)	All hits	Ran-dom hits	All hits	All hits	All hits	All hits	All hits	All hits
1	ind_web-tufs2_2013_1M	421	100	478	100	195	100	234	100	54	54	73	73	94	94
2	ind_wikipedia_2014_1M	587	100	215	100	96	96	253	100	25	25	74	74	47	47
3	ind_web-tufs3_2015_3M	1,251	100	1,368	100	692	100	224	100	277	232	166	166	262	262
4	ind_news crawl_2016_1M	403	100	556	100	141	100	159	100	163	139	79	79	110	110
5	ind-id_web_2017_1M	426	100	366	100	272	100	258	100	91	91	46	46	87	87
6	ind-com_web_2018_1M	533	100	315	100	308	100	343	100	59	59	54	54	96	96
7	ind_news_2019_1M	443	100	463	100	127	100	180	100	121	103	62	62	68	68
8	ind_news_2020_1M	564	100	860	100	236	100	228	100	166	141	188	188	53	53
9	ind_wikipedia_2021_1M	554	100	291	100	140	100	266	100	46	46	106	106	47	47
10	ind_news_2022_1M	408	100	633	100	141	100	192	100	129	110	125	125	56	56
	TOTAL	**5,590**	**1,000**	**5,545**	**1,000**	**2,348**	**996**	**2,337**	**1,000**	**1,131**	**1,000**	**973**	**973**	**920**	**920**

Table 5 shows two total numbers of tokens for each target domain. The left one is all the tokens available in each file. Meanwhile, the right ones show the results of *Random hits* and *All hits* applied in the AntConc, that is ±1.000 tokens in each file, and they are the representative linguistic expressions samples that are ready to be further analyzed. Further analysis using these samples is related to the distinctive metaphors with emotion as the target domain (*ketakutan* and its 6 synonyms). The analysis of the distinctive metaphors is reported in the other article(s).

4. Conclusion

This paper demonstrates a systematic approach for identifying and selecting representative data samples from a corpus containing the Indonesian language files, specifically focusing on emotional target domains. It efficiently sorted and examined files according to the number of sentences and tokens by using the AntConc software to perform a methodical filtering procedure.

The availability of corpus files containing tokens generated from the search of the keywords was conducted through classification and initial filtering. The corpus consisting of 41 files from 20 text sources was classified according to the year and number of sentences. The AntConc software was used to search for emotion words in these files, and the results showed a considerable variation in token counts.

The representative data samples based on the availability of the tokens were selected by applying two phases. First, files with more than 2,000 tokens for specific emotions (*ketakutan, kekhawatiran, kebingungan, kecemasan, keresahan, kepanikan, kegalauan*) were given priority to ensure balanced and representative linguistic expressions. The publication year (2013–2022) and the token count similarity within each year were the selection criteria. Then, to achieve a balanced sample of roughly 1,000 tokens per target domain, a second selection phase that took advantage of random sampling (using AntConc's '*Random hits*' and '*All hits*' settings) was implemented due to the differing token counts. Adjustments were also made when the token counts exceeded 1,000 (for the case of *kebingungan*), ensuring a balanced representation among files.

Acknowledgements

With great appreciation, I, as the first author, would like to thank Udayana University for the financial support in the form of a tuition fee. Additionally, I would like to express my gratitude to my supervisors, Prof. Dr. Ketut Artawa, M.A., Prof. Dr. I Made Rajeg, M.Hum., and Prof. Drs. I Nyoman Udayana, M.Litt., Ph.D for their constant support, and invaluable guidance during this project. I also would like to express my gratitude to Gede Primahadi Wijaya Rajeg, S.S., M.Hum. Ph.D. for providing the statistical analysis help. Their knowledge and support have been crucial to the completion of my study.

References

Anthony, L. (2022). *AntConc (4.2.0)* [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig Corpora Collection: Monolingual Corpora of Standard Size. *Proceedings of the Corpus Linguistics Conference*, 1–13. http://ucrel.lancs.ac.uk/publications/CL2007/paper/190_Paper.pdf

Egbert, J., Larsson, T., & Douglas, B. (2020). *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge University Press. https://doi.org/10.1017/9781108888790

Endarto, I. T. (2020). A Corpus-Based Lexical Analysis of Indonesian English as A New Variety. *Indonesian Journal of Applied Linguistics*, *10*(1), 95–106. https://doi.org/10.17509/ijal.v10i1.24993

Gries, S. T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge. https://doi.org/10.1075/ijcl.17.2.06ruh

Gries, S. Th., & Gales, T. (2024). Talking across the interdisciplinary aisle: A guide for legal and corpus-linguistic scholars and practitioners. *Applied Corpus Linguistics*, *4*(1), 100086. https://doi.org/10.1016/j.acorp.2024.100086

Horvath, A.-S. (2022). ComPara: A corpus linguistics in English of computation in architecture dataset. *Data in Brief*, *42*, 108169. https://doi.org/10.1016/j.dib.2022.108169

Huszka, B. (2020). Metaphors of Anger in Contemporary Bahasa Indonesia: A Preliminary Study. *Journal of Linguistics and Literary Research*, *1*(1), 26–30. https://doi.org/10.32734/lingpoet.v1i1.4694

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Routledge. https://doi.org/10.4324/9781315843674

Lapasau, M., Setiawati, S., Mayasari, I., & Virgana. (2019). Conceptual Metaphors in Modern Indonesian Literature and Their Implication in Language Learning. *Proceedings of the 1st International Conference on Folklore, Language, Education and Exhibition (ICOFLEX 2019)*, *512*, 235–240. https://doi.org/10.2991/assehr.k.201230.044

Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, *9*(12), e22731. https://doi.org/10.1016/j.heliyon.2023.e22731

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. https://doi.org/10.1075/ijcl.18.2.06hun

Mizin, K., & Slavova, L. (2023). The Corpus-based Methodology of Close Emotion Concepts Differentiation: A Case Study of ENVY and JEALOUSY. *Cognitive Studies | Études Cognitives*, *23*. https://doi.org/10.11649/cs.2811

Mulyadi. (2001). Konsep Emosi dalam Bahasa Melayu. *Dewan Bahasa*, *1*(2), 28–35. https://www.researchgate.net/publication/323684446

Oeinada, I. G., Beratha, N. L. S., Sudipa, I. N., & Satyawati, M. S. (2021). *The Use of Japanese Synonymous Verbs GIVE: A Corpus-based Study*. https://doi.org/10.5281/ZENODO.4466802

Oktavianti, I. N., & Sarage, J. (2021). Collocates of "great" and "good" in the Corpus of Contemporary American English and Indonesian EFL textbooks. *Studies in English Language and Education*, *8*(2), 457–478. https://doi.org/10.24815/siele.v8i2.18594

Rajeg, G. P. W. (2019). *Metaphorical Profiles and Near-Synonyms: A Corpus-Based Study of Indonesian Words for HAPPINESS* [Monash University, Australia.]. https://doi.org/10.26180/5cac231a97fb1

Rajeg, I. M. (2013). *Metafora Emosi Bahasa Indonesia* [Udayana University]. https://doi.org/10.5281/zenodo.7604138

Shaver, P. R., Murdaya, U., & Fraley, R. C. (2001). Structure of the Indonesian Emotion Lexicon. *Asian Journal of Social Psychology*, *4*(3), 201–224. https://doi.org/10.1111/1467-839X.00086

Stefanowitsch, A. (2006). Words and Their Metaphors: A Corpus-based Approach. In *Corpus-based Approaches to Metaphor and Metonymy* (pp. 63–105). Mouton de Gruyter. https://doi.org/10.1515/9783110199895

Stoykova, V. (2014). Teaching Corpus Linguistics. *Procedia - Social and Behavioral Sciences*, *143*, 437–441. https://doi.org/10.1016/j.sbspro.2014.07.513

Yosephine. (2022). *Building an Enhanced Resource for Indonesian Sentiment Analysis*. Nanyang Technological University. https://doi.org/10.32657/10356/162004

Yuditha, T. (2013). *Indonesian Metaphorical Conceptualizations of ANGER, LOVE, and HATE: An Overview*. Paper presented at the International Workshop on 'Special Genres' in and around Indonesia, Tokyo, Japan. http://repository.tufs.ac.jp/bitstream/10108/75521/1/B130_123-142.pdf