Jurnal Kata : Penelitian tentang Ilmu Bahasa dan Sastra
Volume 6, Nomor 2, Tahun 2023
E-ISSN: 2502-0706
Open Access: http://publikasi.lldikti10.id/index.php/kata

CORPUS-BASED TERMS EXTRACTION IN LINGUISTICS DOMAIN FOR INDONESIAN LANGUAGE

Wahyu Maulana[1], Eddy Setia[2], Tasnim Lubis[3]

Fakultas Ilmu Budaya, Universitas Sumatera Utara
[1]e-mail: wahyuu.maulana@gmail.com, [2]e-mail: eddy12457@yahoo.com, [3]e-mail: tasnimlubis@usu.ac.id

Article history:

Received
14 April 2022

Received in revised form
6 Agustus 2022

Accepted
12 Agustus 2022

Available online
Oktober 2022

Keywords:
Corpus; Terminology; Term Extraction,

DOI
10.22216/kata.v6i2.908

Abstract

This research aims to extract the mono-lexical and poly-lexical terms from linguistics domain in Indonesian language. As the terminology and lexicology concept is somehow blurry, this research applies CTT by Cabré to do the terms extraction procedure. The corpus-based terminology method is applied in this research to get the best mono-lexical and poly-lexical terms possible. To compile the general and the specialized corpus in this research, AntConc is applied as an instrument. Even though the result is noisy, further analysis about the term limitation manually makes this research semi-automatic. The result shows that the limitation in language and words structure helps this research to delimit the mono-lexical terms extracted in this research. Furthermore, the mono-lexical terms extracted act as the starting point for poly-lexical terms.

INTRODUCTION

Terminology is the study of specialized concepts and their linguistic designations or terms (Faber and Martinez, 2019). Terminology work focuses on the description of domain-specific knowledge structures and how they are transmitted in different communicative contexts. There are many representations about the position of terminology in linguistics as previous researchers debate over the status of terminology as an independent discipline (Dima, 2012). However, the researcher agrees that the result of terminology research has a huge impact on linguistics and any domain it takes its role. This statement is in line with (Faber, 2014) stating that terminology is essential for a wide range of activities, such as technical writing and communication, knowledge acquisition, specialized translation, knowledge resource development, and information retrieval.

The study of terminology is developing into a computer-based study as researchers are overwhelmed with the huge amount of data. For example, Pazienza et al., (2005) do the research about terminology extraction applying linguistics and statistical approach, Peñas et al., (2002) applies corpus method in terminology extraction for information access, and Elfkih & Omri (2012), also does the terminology extraction by applying conditional random fields approach. The similarity between these researches is the term extraction analysis.

While there are several methods for extracting terms, this research, however, is not going to analyze the best method for term extraction. Instead, the researcher is interested in corpus-based term extraction as the corpus-based term extraction is not a fully automated activity even though it is computer-aided. For applying the automated term extraction, the corpus has to apply POS tagging both in general corpus and specialized corpus. This is actually a limitation for Indonesian language as there is no general corpus with POS tagger large enough for this task. The attempt to create a tagged Indonesian corpus is already done

Corresponding author.
E-mail address: wahyuu.maulana@gmail.com

by Fu et al., (2018), Dinakaramani et al., (2014), Christanti et al., (2016) and Kamayani, (2019) with a similar recommendation to expand the available corpus.

The problem for Indonesian language as it has no standardized grammar system by far that leads to confusion for researchers in making the tagset for Indonesia language (Fu *et al.*, 2018). In term extraction, the tagset is useful for extracting the noun or noun phrases in the corpus. As stated by (Pazienza, Pennacchiotti and Zanzotto, 2005), the candidate terms have been mostly identified with noun phrases. Thus, the tagset needed in the corpus is only noun and noun phrase which can be done manually once the term candidates have been extracted.

As terminology is domain-specific knowledge structure, the linguistics domain is chosen for this research. The researcher is going to extract the term in linguistics domain using the corpus-based terminology approach. The corpus-based terminology approach has been done before by Yuliawati et al. (2018) applying the communicative theory of terminology (CTT) and Peñas et al. (2002) applying the corpus-based terminology to information access. The similarity between them is the method of specialized corpus compilation to extract the term candidates by comparing the hits of specialized corpus to the general corpus. Thus, this research creates the linguistics domain corpus from scientific articles as the specialized corpus. Furthermore, the analysis method is adopted from Yuliawati et al. (2018) by applying Mutual Information (MI) score to extract the best term candidates using collocation analysis. The reason to choose the linguistics domain is to extract terms in linguistics domain since there is no similar research in linguistics domain term extraction for Indonesian language.

This research is based on the communicative theory of terminology (CTT) proposed by Cabré. The CTT is a descriptive approach that studies terms and their variants as they appear in texts and envisages the multiple dimensions of specialized knowledge units, as well as their representation and analysis (Faber and Martinez, 2019). Within the CTT, terminological units are regarded as "sets of conditions" derived from a certain knowledge area (Cabré, 2003 in Faber & Rodríguez, 2012). To extract the best term in a specific domain, the terms must be figured in three dimensions; cognitive, linguistic, and communicative.

Yuliawati et al. (2018) applied the CTT analysis on extracting the terms on legal science and administrative science domain resulting in several essential points for the research. The research applies corpus-based terminology research to help the researcher deal with the big quantity of data in the corpora (general and specialized corpus). The general corpus applied for Yuliawati et al. (2018) is the general corpus of social sciences and humanities acting as the reference for legal science and administrative science domain.

The general corpus consists of written and spoken language and it also covers a period of time (Leech, 2002). For instance, the Indonesian corpus in the Leipzig Corpora Collection covers about 13 years. There are also several factors before choosing the general corpus for any research, such as corpus size, genre, varietal difference, and diachrony (Goh, 2011). This research analyzes the general corpus for keyword calculation finding that only genre and diachrony bring significant differences in the numbers of keywords generated. Nelson (2000), states that general corpus is the broadest type of corpus. It is often very large with more than 10 million words containing the variety languages. The examples of the general corpora are The British National Corpus (BNC), the American National Corpus (ANC), and COCA.

The term extraction is distinguished between one-word terms (mono-lexical terms) and multi-word terms (poly-lexical terms) with different extraction methods (Peñas, F and Gonzalo, 2001). The mono-lexical terms are often too polysemic and generic, therefore, it is necessary to provide poly-lexical terms to represent better concepts in a domain (Bourigault and Jacquemin, 1999). As mono-lexical terms are easier to extract by using *keyword* feature, the poly-lexical terms are extracted by using *collocates* feature and further analysis. This is where the CTT takes its place in the research.

To build the corpora for term extraction, there is also a requirement for a specialized corpus in this research. Since the general corpus of Indonesian language is already available online, the specialized corpus for Linguistics domain for Indonesian language is nowhere to be found. The specialized corpus contains texts of a certain type aiming to be representative that can have a small or large amount of data Nelson (2000). This is also a clear difference between a specialized and general corpus. While the general corpus is often very large, the specialized corpus size may vary. Another difference between a specialized and general corpus is the content inside the corpus. Since a general corpus needs to be general from any language source (written or spoken), the specialized corpus contains a specific domain as representative.

The specialized corpus creation is already explained by (Toriida, 2017) focusing on the target materials and word elimination. The target materials in creating the specialized corpus have to consider the context in the corpus and how it will be used. The materials could include a textbook or textbook chapter, graded readers, a collection of scientific articles, course materials, a novel, or a movie script. Another target material mentioned in corpus-based terminology research is doctoral dissertations (Yuliawati, Suhardijanto and Hidayat, 2018) in a certain domain for specialized corpus material. Furthermore, word elimination is done by deleting words from the corpus that are not considered as content words. The word elimination is done by deleting the reference sections and citations, repetitive textbook headings, figure and table headings, proper nouns, and names of institutions or organizations.

Realizing that there is no corpus-based terms extraction for linguistics domain in Indonesian language, the researcher is interested in extracting the terms from the created specialized corpus. As the terms extraction is completed, this research could expand its application onto the other domain and hopefully is able to cover any domain to create the standard for Indonesian language terminology for every domain-specific language.

RESEARCH METHOD

This research is referred to as a corpus-based method for terms extraction. The general corpus is taken from Wortschatz Leipzig corpora collection for Indonesian language. Thus, it underlines that there is no general corpus creation in this research as it is taken from the available general corpus online. Since there is no specialized corpus for linguistics domain for Indonesian language, the specialized corpus is created manually by the researcher.

The source of data for the general corpus is Wortschatz Leipzig corpora collection for Indonesian language (https://wortschatz.uni-leipzig.de/en/download/Indonesian) and it is downloadable. However, there is a limitation for the researcher to download every material provided on the page, thus, the materials taken are from *Mixed* materials with the corpus size of 1,000,000 sentences. For the specialized corpus, the linguistics disertations are taken for the target materials. The disertations using Indonesian language are selected from https://repositori.usu.ac.id/handle/123456789/1082.

There are 62 downloaded files for the specialized corpus in linguistics domain containing 3,302,832 word tokens after the word elimination process. Since the data downloaded are in PDF file format, the data need to be converted into plain text (txt) format using OCR program. The OCR program applied in this research is *AntfileConverter*(Anthony, 2017) as it is a free available software with no page limits to convert PDF files into plain text. After the target materials have been collected, the word elimination process is started. As mentioned before, the words elimination is done by deleting words in the corpus that are not considered as content words (Toriida, 2017). This includes the reference sections, repetitive headings in figures and tables, proper nouns, and names of institutions. In addition, English abstracts are also eliminated from the materials for language uniformity reason. Since the articles are in linguistics domain, the language in the articles may vary from Indonesian, English, or the other local languages. To put the Indonesian language as the scope of the

research, any term candidates (mono-lexical and poly-lexical) in another language besides Indonesian are eliminated from the candidates. Hence, there is also word elimination after the term candidates are collected.

To list the term candidates, *AntConc* (Anthony, 2019) acts as an instrument. For the mono-lexical terms, the keyword list feature is applied to collect the term candidates. The keyword list feature works in an only certain condition; the availability of the general corpus. Thus, the general corpus from Wortschatz Leipzig corpora collection for Indonesian language is added into *AntConc*. After the general corpus is added, the materials for the specialized corpus are then added into the software. The specialized corpus materials are prepared for the word lists in *AntConc* since the keyword list feature needs the word list from the specialized corpus. After the word list is collected, keywords list is created automatically and before putting the keywords into the term list, the words must have the characteristics as follow:

(1) Indonesian language, which is registered in *Kamus Besar Bahasa Indonesia* (KBBI).
(2) Noun (following the mono-lexical terms identity by Pazienza et al. (2005))

Thus, the lists from the keyword feature containing the matched characteristics are defined as the terms for mono-lexical terms.

For poly-lexical terms, the term candidates are taken from the *collocates* feature based on keywords list in *AntConc.* For the collocation settings in *AntConc*, the parameter applied is the *Mutual Information* (MI) score following Yuliawati et al. (2018) and Marzá (2008). The MI score for collocations is 3.00 with the minimal frequency of 5 and a window span of 4L:4R. The window span limit is adopted from Yuliawati et al. since the research is based on Indonesian language. Furthermore, any poly-lexical terms from the collocation are limited by noun phrases only.

The limitations of term candidates for poly-lexical items are based on Marzá (2008) to put more attention in concord analysis by their linguistic forms and the most significant collocates in terms of frequency (mainly in position L3, L2, L1, R1, R2, R3). The concord analysis displayed below the example of the term '*makna'* to '*asali'.*

Table 1. The concord analysis of '*makna'* to '*asali'*

Word	Total	Total Left	Total Right	L3	L2	L1	Center (*makna*)	R1	R2	R3
asali	12	0	12	0	0	0	0	12	0	0

The word '*denotatif*' appears with a frequency of 9 in position R1 without any appearance in the other positions. The example of the analysis is shown below:

"Dari segi	*makna*	*asali,*	*sama sekali tidak ada hubungan…"*
	Center	R1	

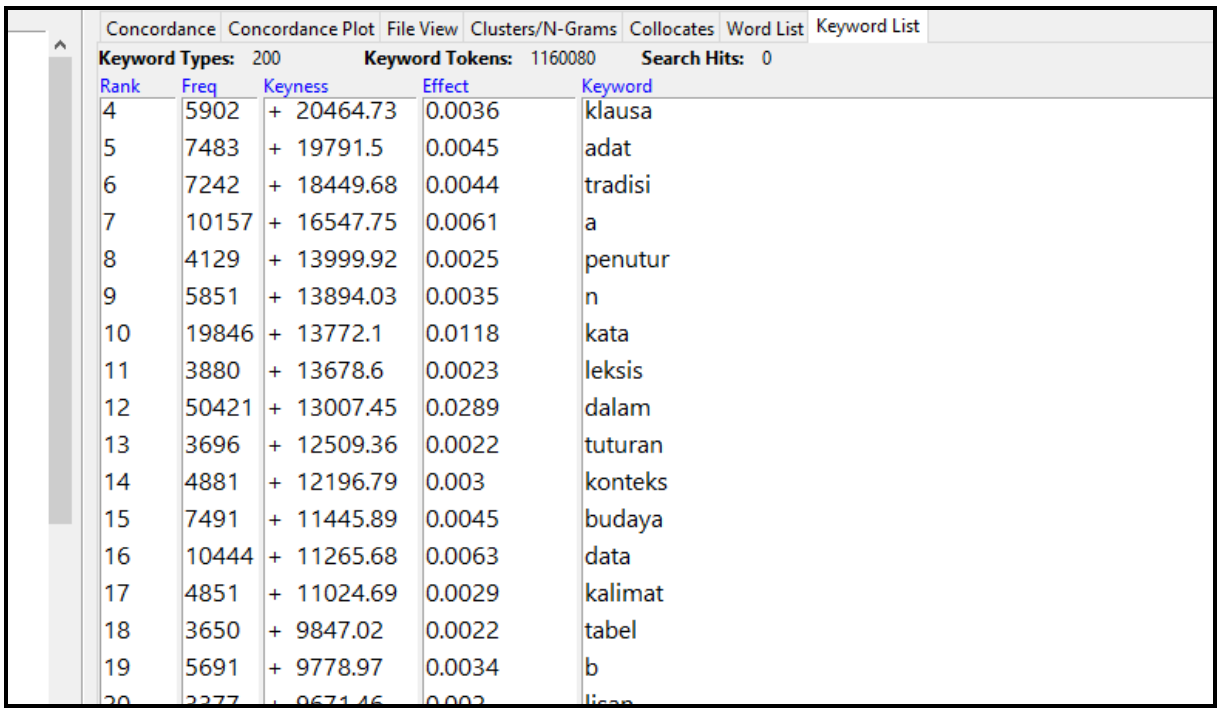
The position of the collocates in the concord analysis shows the most frequent position to where it belongs in the poly-lexical terms to define its exact form based on the corpus. However, this research is capped at only L1 and R1 as the noun phrase in this research is limited by two words only. This also states that the window span in *collocates* feature is set to 1L:1R. Another note for carrying out terminological extraction is the segmentation of the terminological extraction following "dubious to delimit" by Cabré (1993) in Marzá (2008). For this segmentation, (Marzá, 2008) respectfully argues that the steps in term extraction have automatically segmented the terminological units. Thus, the result of the analysis must have the final product of the term extraction.

RESULTS AND DISCUSSION

Since the mono-lexical and poly-lexical terms have different method to extract, the results and discussion is divided into two parts. The first one is the mono-lexical terms extracted from the specialized corpus and the second one is the poly-lexical terms.

a) Mono-lexical Terms

By applying *keyword* feature, the total of mono-lexical term candidates is 1615 word tokens and still considered as term candidates. Before displaying the result of the term extraction, the method for limitation in the data analysis is also displayed for the discussion. After the extraction of the term candidates from the *keyword* feature, there is still some noise in the result. For example, there are several term candidates containing only one letter, such as 'a' (Rank 7), 'n' (Rank 9), and 'b' (Rank 19).



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Keyword Types: 200 **Keyword Tokens:** 1160080 **Search Hits:** 0

Rank	Freq	Keyness	Effect	Keyword
4	5902	+ 20464.73	0.0036	klausa
5	7483	+ 19791.5	0.0045	adat
6	7242	+ 18449.68	0.0044	tradisi
7	10157	+ 16547.75	0.0061	a
8	4129	+ 13999.92	0.0025	penutur
9	5851	+ 13894.03	0.0035	n
10	19846	+ 13772.1	0.0118	kata
11	3880	+ 13678.6	0.0023	leksis
12	50421	+ 13007.45	0.0289	dalam
13	3696	+ 12509.36	0.0022	tuturan
14	4881	+ 12196.79	0.003	konteks
15	7491	+ 11445.89	0.0045	budaya
16	10444	+ 11265.68	0.0063	data
17	4851	+ 11024.69	0.0029	kalimat
18	3650	+ 9847.02	0.0022	tabel
19	5691	+ 9778.97	0.0034	b

Figure 1. Term candidates noise *'a', 'n',* and *'b'* in *AntConc*

Even though the mono-lexical terms are easy to extract, they present a semantic ambiguity and often polysemic (Elfkih and Omri, 2012). So, there is a possibility that these candidates belong in the other domain or as parts of poly-lexical terms that are not being able to stand alone as a mono-lexical term, or perhaps, not a term at all. For making sure that these candidates are not following the characteristics of mono-lexical terms, the *concord* feature is then applied to check whether they are suitable or not. *Concord* highlights the search word in the centre, thus allowing quick detection and analysis of its collocations appearing around it (Marzá, 2008). By analyzing the *concord,* it is found that these candidates containing one letter do not meet the criteria for being term. The *Keyword in Context* (KWIC) in *concord* feature shows that these candidates mainly appear in the phonology or phonetics scientific articles as phonemes.

To check whether they appear in phonology or phonetics articles or not, the *File View* feature comes in handy. It shows the raw text of individual files, allowing the researcher to investigate in more detail the results generated in other features in *AntConc* (Anthony, 2014). Being a phonetic transcription, these letters do not lose their context in linguistics domain but they are not considered as a candidate because as Atkielski (2005) states that phonetic transcription is a written record of the sounds of a spoken language. With that being said,

these letters are the visual representation of a sound and become symbols so that they are excluded from the list. By its scope, this research does not deal with symbols so they are excluded. Moreover, in the *concord* feature, the letters are not always being a phoneme. For instance, the letter *a* can sometimes be a letter list, a variable, or a substitute for other elements in the data (see Figure 2). This also shows that *concord* feature is able to break the ambiguity in the mono-lexical terms to define its context from the natural text.

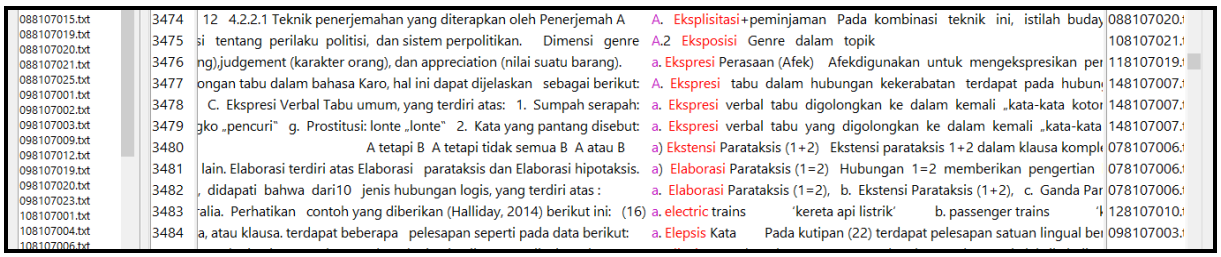
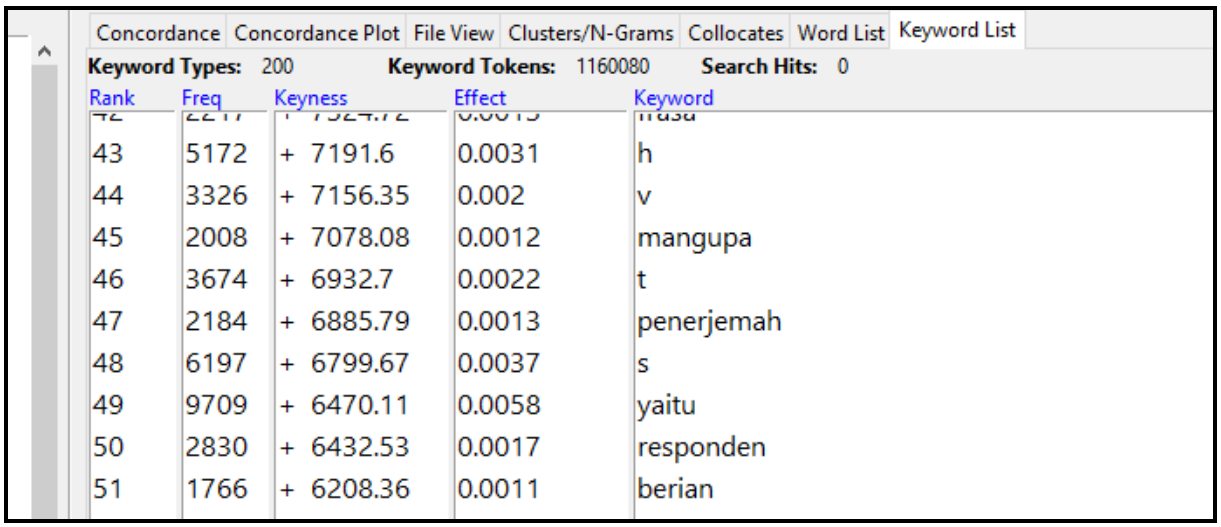


Figure 2. The example of *concord* display for '*a*'

Another characteristic of the mono-lexical terms is the part of speech of each term as a noun. To analyze the part of speech among the candidates, the *concord* feature can also be applied for checking them manually. One of the candidates that is not considered as a noun is shown in Figure 4.3 taken from the *keyword* feature.



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Keyword Types: 200 **Keyword Tokens:** 1160080 **Search Hits:** 0

Rank	Freq	Keyness	Effect	Keyword
43	5172	+ 7191.6	0.0031	h
44	3326	+ 7156.35	0.002	v
45	2008	+ 7078.08	0.0012	mangupa
46	3674	+ 6932.7	0.0022	t
47	2184	+ 6885.79	0.0013	penerjemah
48	6197	+ 6799.67	0.0037	s
49	9709	+ 6470.11	0.0058	yaitu
50	2830	+ 6432.53	0.0017	responden
51	1766	+ 6208.36	0.0011	berian

Figure 3. The term candidate '*yaitu*' (rank 49) in *keyword* feature

To make sure that the word '*yaitu'* (Rank 49) is going to be eliminated from the list, the *concord* feature, once again, comes in handy. The *concord* feature of the word '*yaitu'* is displayed in Figure 4.4 to show that this candidate is eliminated from the list for not being a noun.

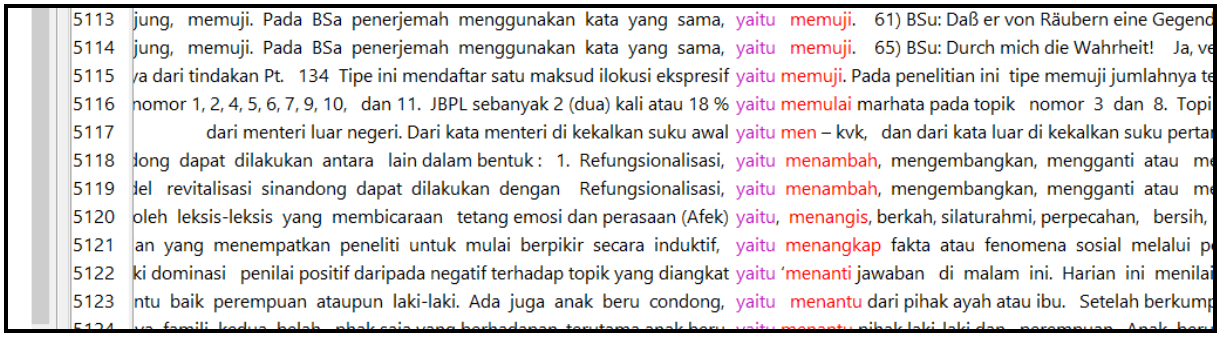


Figure 4. The term candidate '*yaitu*' (rank 49) in *concord* feature

The last characteristic of the mono-lexical candidates is their origin of language. The language origin of each mono-lexical term must be Indonesian language. This characteristic is analyzed by using *Kamus Besar Bahasa Indonesia* (KBBI) by checking them manually. The example of the candidates is the word '*of*' (Rank 118). This candidate is not registered in KBBI as they are foreign words that originated from English.

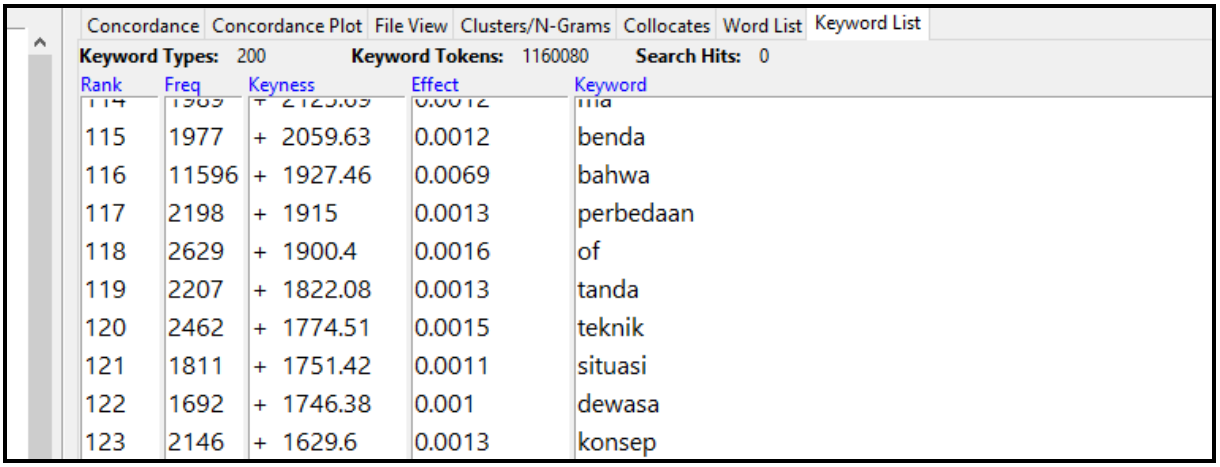


Figure 5. The term candidate '*of*' (Rank 118) in *keyword* feature

After the mono-lexical term candidates are analyzed, the mono-lexical terms are extracted. Table 4.1 displays the 50 highest mono-lexical terms by keyness.

Table 2. The highest 50 mono-lexical terms in linguistics domain extracted from the specialized corpus

Rank	Keyword (mono-lexical term)	Rank	Keyword (mono-lexical term)
1.	bahasa	26.	unsur
2.	teks	27.	linguistik
3.	makna	28.	konsonan
4.	klausa	29.	bunyi
5.	adat	30.	metafora
6.	tradisi	31.	frasa
7.	penutur	32.	penerjemah

Rank	Keyword (mono-lexical term)	Rank	Keyword (mono-lexical term)
8.	tuturan	33.	responden
9.	konteks	34.	berian
10.	budaya	35.	perkawinan
11.	data	36.	nada
12.	kalimat	37.	semantik
13.	tabel	38.	vokal
14.	lisan	39.	kearifan
15.	batak	40.	terjemahan
16.	laki-laki	41.	kajian
17.	struktur	42.	upacara
18.	leksikal	43.	nilai
19.	penerjemahan	44.	sosial
20.	verba	45.	wacana
21.	penelitian	46.	Ciri
22.	fonem	47.	fungsi
23.	analisis	48.	hubungan
24.	melayu	49.	suku
25.	toba	50.	ungkapan

The extracted mono-lexical terms aim to find out which words characterize the text under investigation may be indicative if either what the text is about what is important (Yuliawati, Suhardijanto and Hidayat, 2018). This procedure is suitable to extract the poly-lexical terms, making the mono-lexical terms as head words. Therefore, the extracted mono-lexical terms provide an overview about the main subject in the text. They are regarded as starting point for further analysis in the connection between words.

As (Elfkih and Omri, 2012) stated that the mono-lexical presents semantic ambiguity and sometimes polysemic, the researcher argues that some of the terms somehow fees undergenerated or overgenerated. According to Pasanen (2005), the undergenerated and overgenerated terms can be eliminated to achieve the noise-free term list, however, this procedure comes with its side effect – losing the valid terms. While analyzing the characteristics of the undergenerated and the overgenerated terms in the results, there are no matched characteristics in each term considered as undergenerated (frequency lower than 3) or overgenerated (occurs only once in the source text). Hence, this result shows the valid terms according to the method of the research.

For a reminder, the extracted mono-lexical terms are based on the *keyword* feature in *AntConc* and some manual works by the researcher to have the most accurate result of the mono-lexical terms list. However, it might be impossible to produce a perfect term list automatically or even manually due to the vagueness of the concept *term* itself (Pasanen, 2005). The result of the mono-lexical terms, although considered as not being perfect, is still beneficial for the next step of the research – poly-lexical terms extraction.

b) Poly-lexical Terms

The poly-lexical terms can be extracted from the mono-lexical terms result as it provides an overview of the main subject in the specialized corpus (Yuliawati, Suhardijanto and Hidayat, 2018). The mono-lexical terms result is regarded as the starting point for further analysis, especially in each of their collocation. The collocation analysis is done automatically in *AntConc* by using the *collocates* feature with MI score of 3.00 or higher. Thus, any collocation lower than 3.00 is excluded from the list. After that, the collocation structure is analyzed by using the noun phrase structure in Indonesian language. Lastly, the collocation position in the concord analysis is investigated to check whether the collocation is positioned in L1 or R1. For instance, the analysis of the collocation of the word '*kalimat*' in *AntConc* is displayed.

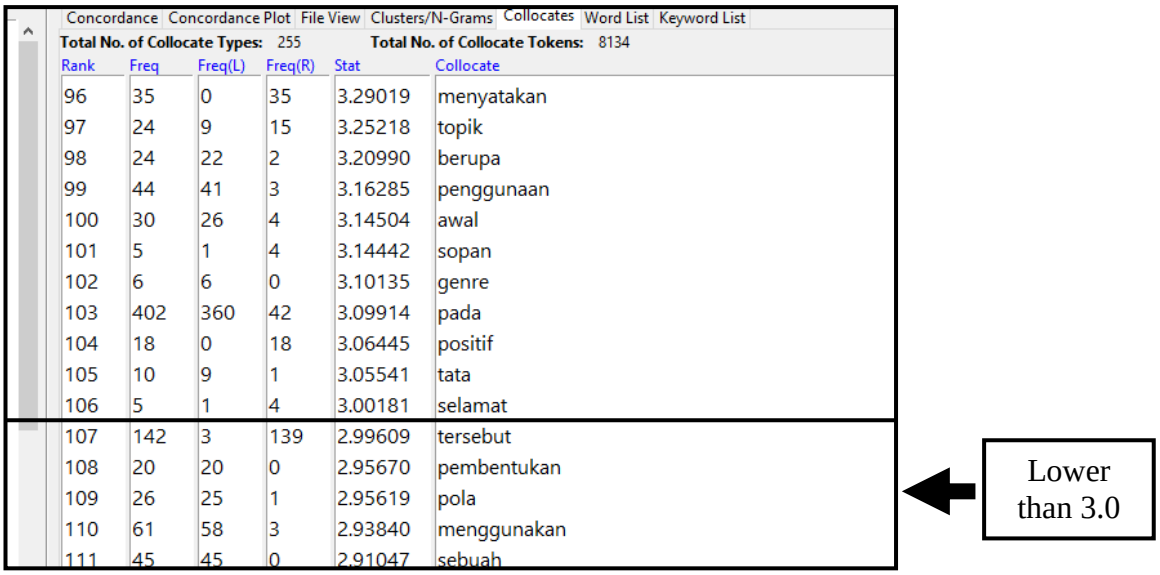


Figure 6. The *collocates* feature applied to word '*kalimat*' in *AntConc*

Figure 4.6 shows that the collocation of the word '*kalimat*' has 225 total collocates with 106 word having a score of 3.00 or more. Hence, the remaining collocates consisting of 119 words are not considered as the candidates. These 106 collocates are analyzed further to get their structure in the *concord* feature. The valid candidates' by their structure is shown below by choosing '*tuturan*'(Rank 51)*, 'koordinatif'* (Rank 11), and *'bersayap'* (Rank 6)as examples.

(1)

"… metafora dalam bentuk	***tuturan***	***kalimat***	***lebih memiliki kekuatan…"***	
	N	N		

(2)

"Perhatikan contoh,	***kalimat***	***koordinatif***	***bahasa Inggris berikut."***
	N	Adj.	

(3)

"Penggunaan	***kalimat***	***bersayap***	***yang dimaksudkan…"***
	N	V	

The candidates must have the noun phrase structure and data (1), (2), and (3) are valid for these characteristics. Moreover, the collocation of the word can be positioned in the left or right. The example of the invalid collocation is shown below.

(4)

"…bentuk-bentuk bahasa	***pada***	***kalimat***	***tersebut mengandung nilai-nilai…"***
	Prep.	N	

The data (4) shows that the word '*pada' (Rank 103)* is not considered as a valid collocation as it does not have the noun phrase structure. As *AntConc* is not built to automatically apply the POS tagging into the words, sometimes, the collocations are noisy as it is mainly based on the frequency without considering the part of speech. This is also the reason why the structure of the collocation is analyzed manually. Another invalid collocation issue comes from the structure of noun + verb as it tends to automatically extract the subject-predicate relationship or two words that are not considered as a phrase in a sentence as Noortyani (2017) states that a phrase is a gramatical unit consisting of two words or more and only occupying a clause element function; subject, predicate, object, complement, and adjunct.

(5)

"… setiap	***kalimat***	***menunjukkan***	***komunitas adat Angkola yang…"***
	N	V	

The data (5) extracts the word '*menunjukkan'* (Rank 88) standing as a predicate. This example, fortunately, is not common in the *collocates* feature and it happens due to the lack of POS tagging preparation before the extraction. The researcher also argues that this issue can be solved by checking the collocation one by one manually after the extraction. However, this might become a massive issue for a large corpus since the researcher could be overwhelmed by the amount of the data. The total word tokens in the specialized corpus in this research is 3,302,832 which is not considered as large data. The amount of the data in a specialized corpus can be varied and there is no claim about the specialized corpus to be as large as possible as it is often created to answer very specific questions (Nelson, 2000).

The next characteristic of poly-lexical terms is their position in the phrase. Any mono-lexical terms are considered as the center of the phrase to their collocation. Since the noun phrase in Indonesian language consists of two words, the collocations included in the extraction are only in position L1 or R1.

(6)

"…dua tuturan	***kalimat***	***direktif***	***dengan kontur deklanasi yang…"***
	Center	R1	

(7)

"Masing-masing	***konstruksi***	***kalimat***	***yang membentuk metafora dapat dilihat…."***
	L1	Center	

The data (6) with using '*direktif*' (Rank 5) is the example of the collocation in position R1 and the data (7) with using '*konstruksi*' (Rank 45) is the example of the

collocation in position L1. Therefore, any collocation positioned in L1 or R1 and considered as a noun phrase is extracted. After all of the analysis, the total of valid collocations of word '*kalimat*' is 55 collocations. Each of the collocations has been analyzed by MI score, noun phrase structure, and position. Table 4.2 displays the 10 highest poly-lexical terms from the word '*kalimat*'.

Table 3. The highest score of the poly-lexical terms of the word '*kalimat*'

No.	Poly-lexical terms	No.	Poly-lexical terms
1.	*kalimat eksklamatif*	6.	*kalimat tanya*
2.	*kalimat syahadat*	7.	*kalimat berdiatesis*
3.	*kalimat tumpuan*	8.	*kalimat koordinatif*
4.	*kalimat direktif*	9.	*kalimat imperatif*
5.	*kalimat bersayap*	10.	*kalimat pasif*

Upon the analysis of the collocation, there is a collocation eliminated from the list – '*BSu'* (Rank 48). While it stands as an abbreviation, this collocation is eliminated as the abbreviation itself is considered a poly-lexical term. Cabré (1998: 86) stated that there are terms that appear to be simple, but upon further examination turn out to be complex and this includes initialisms, acronyms abbreviations and short forms.

The first consideration of '*BSu'* as a part of the collocation of '*kalimat*' comes from the auto-generated collocation in *AntConc*. Further analysis from the *Concord* and *File View* feature in *AntConc* shows that '*BSu*' is the abbreviation of '*bahasa sumber*'. The *Concord* feature shows that there are 781 hits for *'Bsu*' in the data (see figure 7).

Figure 7. The '*BSu*'hits in the *Concord* feature

The whole hits are found in only three files. These files are disertations entitled *TERJEMAHAN UNSUR STILISTIKA TEKS BAHASA JERMAN IPHIGENIE AUF TAURIS KE DALAM BAHASA INDONESIA, TERJEMAHAN SYAIR BAHASA ACEH "MUNAJAT PEREMPUAN SUFI ACEH POCUT DI BEUTONG" DALAM BAHASA INDONESIA:*

ANALISIS STRATEGI PENERJEMAHAN, and *PROTOTIPE MODEL TEKNIK PENERJEMAHAN ISTILAH DAN UNGKAPAN BUDAYA DARI BAHASA INGGRIS KE BAHASA INDONESIA.* The similarity between these disertations are they focus on translation research. The focus or the scope of the study in the disertations can be found in the Keyword section under the Abstract in each of them.Thus, '*BSu*' is considered as an abbreviation with no other variations in the data and is a shortened phrase standing alone as a poly-lexical term.

CONCLUSION

The mono-lexical and poly-lexical terms of linguistics domain in Indonesian language by using corpus method in this research is mainly aims to look for any characteristics of the terms itself. Thus, the corpus-based method is considered a great method to compile and extract the terminology in a specific domain. Although the procedure of the analysis of the mono-lexical and poly-lexical terms are different, the result shows that this is a great start to bring the objective research of terminology in linguistics domain as the term extraction is somehow claimed to be more subjective than objective. With the application of *AntConc* as an instrument, the term extraction can be done semi-automatically. This research also shows that the application of CTT into the term extraction is remarkable as it helps the researcher to do the limitations in the term extraction process to eliminate the noise of the term candidates.

The limitations for the term extraction in this research are stated as follows; a) the specialized corpus must be from a specific domain, b) Each corpus has the same language, c) the part of speech for the terms are noun or noun phrase, d) the MI score for the collocation is set to 3.00, e) the window span for the collocation is 1L:1R, f) the minimum collocation frequency is 5, and g) any terms must be free from any acronyms. By doing the limitations, the extracted terms hopefully reach the most accurate standard for linguistic domain terms. However, the amount of time for the research could be shortened by the presence of proper Indonesian language POS tagging as the syntax analysis of each term can be done automatically by the computer.

By limiting the window span in this research to 1L:1R, it is clear that this research only extracts the 'two words' poly-lexical terms. Expanding the window span in *AntConc* could result in the wider range of poly-lexical terms for the future research. The result of this research can be applied to create the terminology dictionary for linguistics domain in Indonesian language or the bilingual terminology dictionary by developing more corpuses in different language in the data. In this section, the researcher also argues that the presence of semantic prosody and semantic preference in bilingual corpus, mainly for the translation process, could have a huge impact in developing wider terminology dictionary for more than one language.

ACKNOWLEDGEMENTS

The authors would like to thank the Research Institute of Universitas Sumatera Utara for funding this research under TALENTA 2021 Research Grant.

REFERENCES

Anthony, L. (2014) *AntConc (Windows , Macintosh OS X , and Linux) Getting Started (No installation necessary).*

Anthony, L. (2017) 'AntFileConverter'. Tokyo. Available at: https://www.laurenceanthony.net/software.

Anthony, L. (2019) 'AntConc'. Tokyo. Available at: https://www.laurenceanthony.net/software.

Atkielski, A. (2005) 'Using Phonetic Transcription in Class', *Using Phonetic Transcription in Class*, pp. 1–12.

Bourigault, D. and Jacquemin, C. (1999) 'TERM EXTRACTION + TERM CLUSTERING : An Integrated Platform for Computer-Aided Terminology', in *In Proceedings of EACL '99 of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 15–22.

Cabré, M. T. (1998) *Terminology: Theory, methods and applications*. 1st edn. Edited by J. C. Sager. Amsterdam: John Benjamins Publishing Company.

Christanty, V., Pragantha, J. and Victor (2016) 'Part-of-Speech Tagging untuk Bahasa Indonesia Menggunakan Stanford POS-Tagging', in *Seminar Nasional Teknologi Informasi 2016*, pp. 179–185.

Dima, G. (2012) 'A Terminological Approach to Dictionary Entries. A Case Study', in Procedia - *Social and Behavioral Sciences*, pp. 93–98. doi: 10.1016/j.sbspro.2012.10.016.

Dinakaramani, A. *et al.* (2014) 'Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus', *Proceedings of the International Conference on Asian* Language *Processing 2014, IALP 2014*, pp. 66–69. doi: 10.1109/IALP.2014.6973519.

Elfkih, F. and Omri, M. N. (2012) 'A Linguistic Model for Terminology Extraction based Conditional Random Fields', in *ICCRK'2012: International Conference on Computer Related Knowledge*. doi: 10.13140/RG.2.1.3530.7685.

Faber, P. (2014) 'Frames as a framework for terminology', in Kockaert, H. and Steurs, F. (eds) *Handbook of Terminology*. Hardbound. John Benjamins, pp. 14–33. doi: 10.1075/hot.1.02fra1.

Faber, P. and Martinez, S. M. (2019) 'Terminology', *The ASHA Leader*, 24(6), pp. 28–29. doi: 10.1044/leader.ppl.24062019.28.

Faber, P. and Rodríguez, C. I. L. (2012) 'Terminology and specialized language', in *A Cognitive Linguistics View of Terminology and Specialized Language*, pp. 9–32.

Fu, S. *et al.* (2018) 'Towards Indonesian Part-of-Speech Tagging : Corpus and Models', in Yang, E. and Sun, L. (eds) *Proceedings of LREC 2018 Workshop on Belt and Road LRE*. European Language Resources Association (ELRA), pp. 2–7. Available at: http://universaldependencies.org/.

Goh, G.-Y. (2011) 'Choosing a Reference Corpus for Keyword Calculation', *Linguistic Research*, 28(1), pp. 239–256. doi: 10.17250/khisli.28.1.201104.013.

Kamayani, M. (2019) 'Perkembangan Part-of-Speech Tagger Bahasa Indonesia', *Jurnal Linguistik Komputasional (JLK)*, 2(2), p. 34. doi: 10.26418/jlk.v2i2.20.

Leech, G. (2002) 'The Importance of Reference Corpora', *UZEI*, pp. 1–11.

Marzá, N. E. (2008) *The communicative theory of Terminology (CTT) applied to the development of a corpus-based specialised dictionary of the ceramics industry*. Universitat Jaume I. Available at: http://cbueg-mt.iii.com/iii/encore/record/C__Rb5001995__Scorpus dictionary__P0%2C2__Orightresult__X2;jsessionid=191AC477BB19F65A149F8705AE2B575C?lang=cat&suite=def.

Nelson, G. (2000) 'An Introduction to Corpus Linguistics', *Journal of English Linguistics*, 28(2), pp. 193–196. doi: 10.1177/00754240022004965.

Noortyani, R. (2017) *Buku Ajar Sintaksis*. 1st edn. Edited by M. Arsyad. Yogyakarta: Penebar Pustaka Media. Available at: https://scholar.google.co.id/scholar?hl=id&as_sdt=0%2C5&q=jurnal+artikel+ilmiah&btnG=.

Pasanen, P. (2005) 'A Term List or a Noise List? How Helpful is Term Extraction Software

when Finnish Terms are Concerned?', in Madsen, B. N. and Thomsen, H. E. (eds) *7th International Conference on Terminology and Knowledge Engineering*, pp. 375–384.

Pazienza, M. T., Pennacchiotti, M. and Zanzotto, F. M. (2005) 'Terminology extraction: An analysis of linguistic and statistical approaches', *Studies in Fuzziness and Soft Computing*, 185(2005), pp. 255–279. doi: 10.1007/3-540-32394-5_20.

Peñas, A., F, V. and Gonzalo, J. (2001) 'Corpus-based terminology extraction applied to information access', *Proceedings of Corpus Linguistics 2001*, (August 2013).

Peñas, A., Verdejo, F. and Gonzalo, J. (2002) 'Terminology Retrieval: Towards a synergy between thesaurus and free text searching', *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2527(Hertzberg), pp. 684–693. doi: 10.1007/3-540-36131-6_70.

Toriida, M.-C. (2017) 'Steps for Creating a Specialized Corpus and Developing an Annotated Frequency-Based Vocabulary List', *TESL Canada Journal*, 34(1), pp. 87–105. doi: 10.18806/tesl.v34i1.1257.

Yuliawati, S., Suhardijanto, T. and Hidayat, R. S. (2018) 'A Corpus-based Analysis of the Terminology of the Social Sciences and Humanities', in *IOP Conference Series: Earth and Environmental Science*. IOP Publishing. doi: 10.1088/1755-1315/175/1/012109.