JOURNAL LA SOCIALE

VOL. 06, ISSUE 01 (103-110), 2025
DOI:10.37899/journal-la-sociale.v6i1.1762

A Linguistic Corpus-Based Analysis of the Synonym Differences

Minhui Zhang[1], Teguh Setiawan[1]

[1]*Universitas Negeri Yogyakarta*

**Corresponding Author: Minhui Zhang*
E-mail: minhui.2022@student.uny.ac.id

Article Info

Article history:
Received 8 November 2024
Received in revised form 11 January 2025
Accepted 20 January 2025

Keywords:
Synonyms
Corpus
Collocations
Semantic Prosody

Abstract

In learning vocabulary, synonyms become a big difficulty for Indonesian language learners as a second language, because they have the same or similar semantic meanings. The emergence of the corpus has brought about new research methods for the identification and analysis of synonyms. This study aims to compare semantic collocations and prosody of Indonesian synonyms 'secepat' and 'so that' so as to find differences in usage between the two words. This research method uses qualitative and quantitative methods. The results of this study are: the frequency of 'secepat' is higher than 'soleh' in the Indonesian Web Corpus (IndonesiaWeC), and 'soleh' is only side by side with the class of prepositions. Meanwhile, words that are collocated 'so' tend to be personal pronouns. There are 6 similar collocations between 'so' and 'so that'. Meanwhile, 'so that' and 'so that' have the same semantic prosody that is neutral.

Introduction

Vocabulary is an important aspect in second language learning and plays a vital role. How much vocabulary a learner masters will affect the ability to acquire a second language. Wilkins in (Wang, 2019) said that only a little information can be conveyed without grammar, while no information can be transferred without vocabulary. When learning vocabulary, we should not only pay attention to the vocabulary itself, but also consider collocation. For example, the word 'work' often collocates with 'same' but 'with' even though the words 'same' and 'with' have the same semantic meaning. Knowing collocation can help learners master the use of vocabulary (Karoly, 2005; Bui, 2021).

In vocabulary learning, synonyms are a big difficulty for second language learners because they have the same or similar semantic meanings. For example, 'want' and 'want'. For beginners learning Indonesian, they will most likely make sentences like 'Aldi wants to fall when he can't control his bike' but 'Aldi wants to fall when he can't control his bike'. Both words have the same meaning but when used in certain contexts it will cause an unusual meaning. 'Want' comes from oneself, while 'want' is something that will happen. Aldi will fall off his bike not from himself he wants to fall. Therefore, even though synonyms have the same or similar meanings, their use will be different. When synonyms are juxtaposed with certain words in certain contexts, they will show different semantic prosody as well. (Lan & Wang, 2021) once studied the synonyms improve and enhance, the results are as follows: improve is often juxtaposed with significantly, greatly, steadily which are positive, so the word improve is positive. While enhance can not only be paired with significantly, greatly, but also can be paired with thus, thereby which are neutral, then the word enhance can be positive and neutral. Therefore, how to distinguish and choose the right synonyms to meet the needs of a particular context is always important in the process of learning a second language and these things are not easy for students (Ma, 2022; Faemthaisong & Khamkhien, 2021).

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

Synonyms in KBBI are word forms that have similar or the same meaning as another language (Fitrah & Afria, 2024; Priasty, 2024). According to (Zhang, 2022) synonyms are expressions such as words, phrases or sentences that have more or less the same meaning as other expressions. The use of synonyms cannot be avoided both in verbal communication and writing activities. Especially in writing activities, if the writer uses a vocabulary continuously, the reader will be bored to continue reading. The use of synonyms can not only enrich vocabulary, but is also a form of second language skills for learners (Hang & Du, 2021; Fazliddionovna, 2023).

In traditional Indonesian language teaching in China, teachers always compare the differences in semantic meaning in Mandarin to Indonesian synonyms, then provide examples of synonym sentences (Yolanda & Setyono, 2023). In a way that is out of context, it will actually confuse students in the use of synonyms, and cause inappropriate use. Finally, students cannot master the use and differences of synonyms well, as a result they cannot improve their Indonesian. (Polizzi et al., 2024) said that the science of semantic prosody is very valuable for distinguishing synonyms and choosing translation equivalents for language learners, but semantic prosody debates are also difficult to observe through direct feelings for speakers of both languages. So, it takes a lot of data to support it. The emergence of the corpus has brought new research methods for identifying and analyzing synonyms. (Siregar et al., 2024) said that a large-scale corpus provides a new perspective and a new way to distinguish the meaning of synonyms, namely studying the meaning of synonymous words through natural contexts. Then, learners can feel the semantic prosody of synonyms in certain contexts so that they can.

Several studies related to corpus-based synonym analysis include (Puspita, 2016) who discussed the meaning of synonyms 'mau', 'ingin', 'hendak' and 'akan' by utilizing the Leipzig corpus and Sketch engine. There are results that 'mau' is used more in informal contexts, while 'ingin' is used in more formal contexts. 'Akan' often collocates with 'denial', 'possibility' and 'certainty'. 'Hendak' collocates with 'time adverbs', 'assumptions', and is widely used in Malay. The second is by (Ermanto et al., 2023) who studied the frequency, collocation, register and semantic prosody of the synonyms every and each using the COCA corpus. The results of this study are that the frequency of every and each is almost the same, but each is more formal than ever. Both words often collocate with nouns, while rarely with verbs or adverbs. Every and each both have neutral semantic prosody. The third is by Liu (2021) which is about the frequency and collocation of synonyms improve and promote using COCA. This study shows that the frequency of improve is higher than promote in the COCA corpus. Improve often collocates with words related to human life, such as quality, health, performance, skills and others. While promote often collocates with words of macroscopic significance, such as development, growth, efforts and others. The last is (Gu, 2017) research on the collocation and semantic prosody of synonyms 'cause' and 'result' using the Sketch Engine corpus. The results of the study are as follows: 'cause' and 'result' have similar collocation behavior, but have their own preferential collocations as well. For example, 'cause' only coexists with 'pollution', 'congestion' and others. While 'result' only coexists with 'irritation', 'injury', 'blindness' and so on. The semantic proposition between 'cause' and 'result in' is negative.

There have been many studies that examine synonyms based on the corpus as described above, but the words studied still focus on adjectives such as every, each and verbs such as improve, promote, 'menyebabkan', 'mengakibatkan'. While the focus on prepositions is still very lacking. Therefore, the researcher tries to analyze the semantic prosody of the prepositions 'sehingga' and 'supaya' which are synonymous based on their collocations juxtaposed on the right which are shown in the data source taken from the Indonesian Web Corpus (IndonesiaWaC). This study is expected to help Indonesian language learners better understand and differentiate the

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

use of 'sehingga' and 'supaya' which have similar meanings through their collocation behavior and semantic prosody. Basically, the purpose of this study is to provide deeper information about a set of synonymous words, namely 'sehingga' and 'supaya' so that it is useful for Indonesian language students, teachers or researchers in the field of teaching writing or compiling Indonesian and Mandarin dictionaries.

Methods

This study uses qualitative and quantitative methods because the data taken from the corpus needs to be interpreted further and in detail by researchers using qualitative methods and the aim is to understand the meaning behind the data so that they can find out the semantic prosody between the words 'sehingga' and 'supaya' (Triadi & Nur, 2024). The quantitative method is reflected in the process of collecting and organizing data that utilizes Sketch Engine which is a multifunctional platform. To analyze the data, researchers use word sketches and concordances in the Sketch Engine corpus. The data analysis technique in this study is the contrastive analysis technique. Data was obtained from the Indonesian Web Corpus (IndonesiaWeC) which consists of more than 100 million words. In this study, 'sehingga' and 'supaya' which have the same word class will be selected as research objects.

Results and Discussion

According to KBBI, 'sehingga' is a conjunction to mark the effect and 'supaya' is a conjunction to mark the goal or hope; hopefully it reaches its intended purpose; so that. Based on this definition, it seems that 'sehingga' and 'supaya' can be considered synonyms because they have the same meaning, namely words to mark the purpose of an event or action. Apparently, it is clear that Indonesian language learners will be confused when choosing the words 'sehingga' and 'supaya' because the definition of these two synonymous words in KBBI does not show their collocation and prosodic semantic characteristics.

'Sehingga' appears 105,000 times in the Indonesian Web Corpus (Indonesia WaC), which contains 90,120,046 words, while 'supaya' appears 23,042 times. The *mutual information* (MI) value is often used in collocation research to measure the strength of the relationship between words x and y. In the popular software used to process corpus data, *Sketch Engine* also has a term called *typicality score.* Here, a higher value is more suitable for the collocation of words x and y. The following table shows the colloquialisms on the right followed by 'sehingga' and 'supaya'. The important data from the Indonesian Web Corpus (Indonesia WaC) are filtered by *typicality scores* that are more than 5. Then, the data is sorted from high to low. After sorting, the data is in accordance with the previous standard, there are 88 colloquialisms related to 'sehingga' and 26 colloquialisms related to 'supaya'.

Table 1. Colloquial words 'so that' and 'so that' with Typicality Score > 5 Indonesian Web Corpus (IndonesiaWaC)

so that			**so that**		
collocate	**frequency**	**Score > 5**	**collocate**	**frequency**	**Score > 5**
can	3321	8.2	You	793	8.4
He	2341	8.1	Don't	423	8.0
they	3161	7.7	they	1195	6.6
No	4885	7.6	He	483	6.4
cause	663	7.5	you	104	6.3
capable	851	7.5	can	666	6.2
now	734	7.4	Can	545	6.1
cause	622	7.3	We	812	6.1

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

Can	1542	7.2	No	1477	6.0
he	1127	7.2	he	297	5.8
Finally	572	7.0	always	45	5.8
make	702	6.9	every	174	5.8
We	1610	6.8	I	200	5.7
happen	666	6.8	still	135	5.7
become	1318	6.7	it is	34	5.4
allow	372	6.7	looks	37	5.3
need	579	6.7	all	179	5.3
not	677	6.6	you	50	5.3
Lots	775	6.5	more	335	5.3
You	768	6.5	easy	67	5.2
when	537	6.5	avoid	27	5.2
easy	377	6.4	obey	26	5.2
difficult	331	6.4	quick	56	5.1
every	489	6.4	we	137	5.1
expected	290	6.3	man	152	5.1
produce	298	6.3	people	82	5.0
must	664	6.3			
result in	259	6.2			
for	593	6.2			
if	295	6.1			
we	438	6.1			
I	431	6.1			
make it easier	223	6.1			
to form	249	6.1			
arise	232	6.0			
all	445	6.0			
when	265	5.9			
I	777	5.9			
public	421	5.9			
he	257	5.9			
often	264	5.9			
If	346	5.8			
appear	235	5.8			
if	302	5.8			
will	869	5.7			
very	407	5.7			
person	663	5.7			
on	963	5.7			
created	163	5.6			
required	184	5.6			
Now	246	5.6			
more	522	5.5			
There is	597	5.5			
reduce	157	5.4			
later	143	5.4			
give birth to	150	5.4			

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

process	206	5.4			
almost	170	5.4			
to	539	5.4			
all over	207	5.4			
You	201	5.4			
only	349	5.4			
amount	180	5.4			
day	260	5.3			
forced	134	5.3			
it happened	123	5.2			
the more	177	5.2			
student	143	5.2			
child	234	5.2			
matter	251	5.2			
make	135	5.2			
push	128	5.2			
give	189	5.1			
moment	244	5.1			
formed	117	5.1			
price	141	5.1			
need	122	5.1			
in a way	299	5.1			
impressed	111	5.1			
obtained	117	5.1			
results	159	5.0			
reasonable	109	5.0			
not enough	142	5.0			
increase	128	5.0			
looks	116	5.0			
own	203	5.0			
you	119	5.0			
What	210	5.0			

In the table above, it can be seen that the collocation of the word on the right that follows the word 'so that' with the highest frequency is 'tidak', 'dapat', 'they', 'ia', 'kita', 'bisa', while 'supaya' often collocates with 'tidak', 'they', 'kita', 'kamu', 'dapat'. It can be concluded from the collocation that 'so that' and 'supaya' which are synonyms also have almost the same collocation behavior. In addition, the data also shows that 'so that' and 'supaya' also have their own preferences. For example, 'so that' only goes hand in hand with 'ke', 'pada' which are prepositional words. Meanwhile, the words that are collocated with 'supaya' tend to be with personal pronouns, such as 'kamu', 'aku', 'kami', 'Kalian'.

When sorted by frequency, it can be seen that 'sehingga' and 'supaya' have very similar collocations in all selected high-frequency collocations. These synonyms can share 6 collocations. The following table shows the collocations shared by 'sehingga' and 'supaya' and grouped according to category.

Table 2. Collocate Categories Shared by 'So' and 'So'

Category	**Collocate**
adverb	can not

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

verb	Can
personal pronoun	they, he, we

Of the 88 most frequently used words on the right, the word 'so that' appears 67 times more with neutral mood, such as 'can', 'ia', 'they', 'become', 'when', and 'many'. Therefore, it is clear that the word 'so that' collocates more often with neutral words. The following table shows the distribution of the nature of the word 'so that'.

Table 3. Distribution of the Characteristics of the Colloquial Word 'so that' (Typicality Score > 5)

Positive (13)	able, can, expected, produce, facilitate, reduce, more, give birth to, increasingly, encourage, give results, increase
Negative (8)	not, to conclude, to cause, to be difficult, to result, to arise, to be forced, to be lacking
Neutral (67)	can, he, they, now, she, finally, make, we, happen, become, possible, need, not, many, you, when, easy, every, must, the, if, we, I, form, all, when, I, society, he, often, if, appear, if, will, very, people, on, created, needed, now, there, later, process, almost, to, all, you, only, number, day, happen, students, children, things, make, when, formed, price, need, in, impressed, obtained, reasonable, seem, have, you, what

An example of a sentence with a concordance search is the following: 1) The conclusion states that everyone will naturally be able to understand the signals given by their partner, so there is no need to learn to understand them; 2) Erase the squares and small box lines so that the map of Australia is visible; 3) On the contrary, he must wait for the time when he is called by God, so that he has a testimony about his calling, so that he feels certain and sure that his calling comes from God; 4) The system uses plasma lasers placed where people normally cross and projects the shadow of the person crossing onto a wall of light so that it can be directly seen by drivers; 5) The neighbors knew of Fina's suffering, but her wounds smelled so bad that they died; 6) The Khilafah State must create attractive trade terms so that we can obtain North Korean agricultural equipment and also benefit from their agricultural techniques; 7) News of this war spread widely throughout Eastern Europe and was also heard by the Byzantine Emperor, so the Kuffar soldiers were happy and sent letters of thanks and joy to Timurlane.

Table 4. Distribution of the Characteristics of the Colloquial Word 'supaya' (Typicality Score > 5)

Positive (2)	can, more
Negative (4)	never, never, always, avoid
Neutral (20)	you, they, he, you, can, we, he, every, I, still, it, looks, all, you, easily, obey, immediately, we, humans, people

By examining the word on the right, it can be seen that there are 20 out of 26 words with neutral mood in the collocation of 'supaya' that are used, for example 'you', 'they', 'he', 'you', 'can', and 'we'. Therefore, it is clear that the word 'supaya' more often collocates with neutral words so that it has a neutral semantic prosody. An example of a sentence with a concordance search is the following: 1) Actually, at that time the soles of my feet were still intact but they were pale white, there was no blood flow and according to the doctor, they had to be amputated immediately so as not to damage other parts; 2) The shape of this bottle was designed so that it could be criticized worldwide, even if it was broken, and has been marked as a trade mark; 3) The palace was then repaired following the model of a medieval palace so that it looked

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

good from the outside; 4) However, because the place was considered unsafe, the transaction location was moved to avoid police officers; 5) Through the political parties they also straddle, they try to maintain a role in all state institutions.Also, they will demand that the government be more transparent and want every government activity to be reported.

Despite the fact that the explanation of KBBI V about 'sehingga' and 'supaya' does not show a neutral semantic prosody, it is clear that these two words are often used in a neutral context, so that their semantic prosody tends to be neutral. This also supports Phoocharoesil's (2010) opinion that corpus-based information data often has more information than the information available in the dictionary because of the grammatical and collocational patterns associated with more accurate words.

Conclusion

Through a corpus that has a large amount of language data, second language learners can find out the frequency of word occurrence, word collocation with other words, and how the semantic atmosphere displayed by a word through its concordance search in the corpus so that they can distinguish synonyms accurately. The method of distinguishing synonyms based on the corpus gives us a new perspective to observe language behavior, and also shows that based on detailed evidence provided by the corpus, we can describe the nuances of synonyms scientifically and comprehensively. As synonyms, 'sehinga' and 'supaya' have almost the same meaning in the dictionary, but after observation and analysis in the corpus, it can be concluded that the frequency of 'sehinga' is higher than 'supaya' in the Indonesian Web Corpus (IndonesiaWeC), and 'sehingga' is only side by side with the class of prepositional words. Meanwhile, the words that are collocated with 'supaya' tend to be with personal pronouns. There are 6 collocations that are the same between 'sehinga' and 'supaya'. Meanwhile, 'sehingga' and 'supaya' have the same semantic prosody, namely neutral.

References

Bui, T. L. (2021). The role of collocations in the English teaching and learning. *International Journal of TESOL & Education*, *1*(2), 99-109. http://dx.doi.org/10.11250/ijte.01.02.006

Ermanto, S. P., Ardi, H., & Juita, N. (2023). *Linguistik Korpus: Aplikasi Digital Untuk Kajian Dan Pembelajaran Humaniora*. PT. RajaGrafindo Persada-Rajawali Pers.

Faemthaisong, P. I. M. P. A. T. I. P. A. R. N., & Khamkhien, A. (2021). *A corpus-based study of English synonyms: General, common, and typical* (Doctoral dissertation, Doctoral dissertation, Thammasat University).

Fazliddionovna, Z. M. (2023). Synonyms In Russian Philology (Literature Review). *International Journal of Advance Scientific Research*, *3*(11), 19-26. https://doi.org/10.37547/ijasr-03-11-05

Fitrah, Y., & Afria, R. (2024). The Analysis of Synonym Relation Meaning in Kerinci Language: A Semantic Study. *Journal of Languages and Language Teaching*, *12*(1), 404-415. https://doi.org/10.33394/jollt.v12i1.9415

Gu, B. J. (2017). Corpus-based study of two synonyms: obtain and gain. *Sino-US English Teaching*, *14*(8), 511–522. http://dx.doi.org/10.17265/1539-8072/2017.08.006

Hang, P. T., & Du, N. T. (2021). High school students' attitudes towards the use of a synonym and antonym dictionary in learning vocabulary. *TNU Journal of Science and Technology*, *226*(13), 109-117. http://dx.doi.org/10.34238/tnu-jst.5034

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)

Karoly, A. (2005). The importance of raising collocational awareness in the vocabulary development of intermediate level learners of English. *Eger Journal of English Studies*, (5.), 58-69.

Lan, X., & Wang, Y. (2021). The Usage of Two Chinese Synonyms for Learners: A Corpus-Based Approach. *2021 2nd Asia-Pacific Conference on Image Processing, Electronics and Computers*, 590–593. https://doi.org/10.1145/3452446.3452590

Ma, Z. Z. Z. (2022). On the identification of near-synonyms in teaching Chinese as a foreign language. *Frontiers in Educational Research*, *5*(7.0). https://doi.org/10.25236/FER.2022.050715.

Polizzi, D., Bernardini, S., & Ferraresi, A. (2024). Evaluation in a cross-linguistic perspective: Investigating semantic prosody across English and German near-synonyms. *Journal of Corpora and Discourse Studies*, *7*. https://doi.org/10.18573/jcads.120

Priasty, I. (2024). *The students' antonyms and synonyms mastery of the grade vii MTs Ell-Firdaus Cikampak-Labuhanbatu Selatan* (Doctoral dissertation, UIN Syekh Ali Hasan Ahmad Addary Padangsidimpuan).

Puspita, D. (2016). Pemanfaatan Korpus dalam Analisis Makna Kata Bersinonim mau, ingin, hendak dan akan. *Bahasa-Bahasa Daerah Di Indonesia*.

Siregar, E. D., Sinar, T. S., & Prihantoro, P. (2024). Turunan Verba "Juang" Berbasis Korpus serta Bandingannya dengan KBBI Edisi VI. *Kajian Linguistik Dan Sastra*, *3*(3), 288–304. https://doi.org/10.22437/kalistra.v3i3.34452

Triadi, R. B., & Nur, A. M. (2024). *Metode Penelitian Bahasa*. Langgam Pustaka.

Wang, Q. (2019). A Corpus-based Contrastive Study on Semantic Prosody of English near Synonyms: A Case Study of Motive and MotivationA Corpus-based Contrastive Study on Semantic Prosody of English near Synonyms: A Case Study of Motive and Motivation. *Journal of Arts and Humanities*, *8*(1), 1–15. https://doi.org/10.18533/journal.v8i1.1552

Yolanda, Y., & Setyono, B. (2023). Why do Words with Negative Connotations Still Exist? A Corpus-Based Analysis of the Words 'Handicapped','Diffable', and 'Disability'. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4). https://doi.org/10.21659/rupkatha.v15n4.15

Zhang, L. (2022). Studi Berbasis Korpus: Perbandingan Kolokasi Dan Prosodi Semantik Sinonim Bahasa Indonesia "Menyebabkan" Dan "Mengakibatkan." *Mabasan*, *16*(1), 153–176. https://doi.org/10.62107/mab.v16i1.517

ISSN 2721-0960 (Print), ISSN 2721-0847 (online)