A Corpus-Driven Approach on Learning Near Synonyms of Pain in Indonesian

Haniva Yunita Leo
Australian National University, Australia

Abstract

Pain is human-universal since it is experienced by people across the world. However, since it is related to personal feelings, different people may feel it in a different way and rely on language to communicate. This paper presents a cross-cultural comparison of the study of the emotion of pain in Indonesian by examining the usage of two near-synonyms: *sakit* and *nyeri*. This study aims to provide a new insight for L2 learners of Indonesian regarding the study of emotion. A corpus-driven method by using the usage-feature analysis (Glynn, 2010b) is employed to test the hypothesis on the semasiological structure of pain from Indonesian dictionary. The corpus data of Indonesian News 2020 with a total of 15,206,710 tokens were extracted from the Leipzig Corpora Data Collection of Indonesian (Goldhahn et al., 2012). A total of 400 examples of *sakit* and *nyeri* were extracted from the corpus data using AntConc version 4.1.2 (Laurence, 2022) for manual annotation. The manual coding of the lexemes was conducted based on cross-linguistic dimensions of pain proposed by Wierzbicka (2016). After manual annotation, two statistical analyses were conducted in R (R Core Team, 2022), namely Binary Correspondence Analysis (Glynn, 2014) and Binomial Regression Analysis (Levshina, 2015). The result of exploratory analysis shows that *sakit* and *nyeri* can be distinguished by bodily focus and intensity. However, the confirmatory analysis confirms bodily focus as the significant predictor. It means *nyeri* is strongly associated with pain on the part of body relative to *sakit.* The finding of the current study may have an implication for the possibility of combining cross-cultural competence with L2 vocabulary learning by making use of corpora in L2 learning design.

Keywords: corpus-driven, learning, near-synonyms, pain, Indonesian

The study of near-synonyms has been an interesting topic in language learning these days. Since the use of corpora in language study increases, it is found that synonymous words are not actually synonymous when observed by their use (Reppen & Simpson-Vlach, 2020). Reppen and Simpson-Vlach (2020) elaborated on this claim by giving instances of how dictionary lists "resulting copula" *become*, *turn*, *go*, and *come* as synonyms. However, when these words are researched in corpora, they might differ by their context of use. Therefore, the term "near-synonym" is used instead of synonyms because they represent a group of words with closely similar meanings (Yu et al., 2016). This subtle difference between near-synonyms can be revealed by using corpus as a tool, as it may enrich the exploration of word senses that have been provided by dictionary for the purpose of foreign language learning (Reppen & Simpson-Vlach, 2020).

Among the near-synonyms that have been studied by scholars (Krawczak, 2015; Rajeg , 2016; Rajeg et al., 2020; Yu et al., 2016), the social emotion of pain becomes increasingly interesting due to its variation of meaning across cultures. This has been highlighted by Priestley (2016), who examined words and expressions of pain used by Koromu speakers. In her study, she described how the influences of culture, environment, and key life events have shaped the concept of pain in the Koromu language, making the terms vary in different constructions. Moreover, Wierzbicka (2016) also emphasized how pain is related to personal feelings, hence different people may feel it in a different way and rely on language to communicate what they feel about it.

As language becomes the only tool for expressing pain, it sometimes turns problematic when people from different cultures communicate, especially in medical practice. Kalengayi et al. (2012) described challenges faced by health practitioners in cross-cultural care. In their study related to problems in communication between international migrants and caregivers in Sweden, they proposed the translation of key documents and trained interpreters to cope with the challenges in communication. While translation and interpretation become the solution, it should be noted that not all languages have the same concept of pain as it is expressed in English (Wierzbicka, 2016). Thus, studying it cross-culturally will shed light on the cultural concept of pain from different languages to find the nearest equivalence to deal with translation issues and develop cross-cultural communicative competence of L2 learners of Indonesian.

Byram and Wagner (2018) highlighted the importance of including cross-cultural knowledge in language education and encouraged it to be part of language learning. They assumed that having cross-cultural communicative competence will enable learners "to engage in intercultural communication, to think and act critically, and to negotiate the complexities of today's world" (Byram and Wagner, 2018, p. 141). In line with their statement, Stone (2023) argued that utilizing a dictionary and textbook would never be adequate to learn and comprehend a foreign language because learners will not be able to communicate and negotiate successfully in a real-world communication as they lack the context of use. Therefore, developing cross-cultural communicative competence in language learning will help learners engage with the cultures where the language is spoken.

Despite the importance of the cross-cultural study on pain from the sense of language learning, Wierzbicka (2012) argued that many studies only focus on the medical and social aspects of pain and tend to ignore the cross-cultural and cross-linguistic perspectives. While this area is less explored, it is essential to note that pain is why people seek medical care, and language is used to communicate it. As language is used for communicating pain, previous studies on medical communication (Dorsey et al., 2022; Unger et al., 2022) have shown ineffective communication experienced by patients from different cultural backgrounds due to language barriers. For example, a Korean immigrant in Australia explained her struggle to describe the symptom she felt related to pain in her stomach (Yoon, 2007). It, therefore, signals the need for cross-cultural analysis to shed light on L2 learning.

Based on this crucial issue in language learning, the present study provides a cross-cultural comparison of the study on the emotion "pain" in Indonesian by examining the usage of two lexemes, *sakit* and *nyeri*, from a corpus and comparing their patterns of use quantitatively. While there have been few previous corpus studies focusing on pain in Indonesian, this study may benefit the study of near-synonyms in two ways. First, for L2 learners of Indonesia, this study presents a new insight into the use of *sakit* and *nyeri* and how the quantitative approach may help them understand to what extent these words may differ significantly or insignificantly in their context, in which sakit is associated with pain in the body as a whole that includes both of physical and emotional phenomena, while nyeri is more localized and physical. Second, this study may enrich the cross-cultural concept of pain in Indonesian so that translators may compare it with other languages to choose the appropriate equivalence of pain in Indonesian.

This paper is organized into several sections. It will begin with the theoretical framework related to this study, followed by the research methods. Then, in the results section, it presents two statistical analyses to discriminate near-synonyms of pain as well as discussion. In the last section, recommendations are made before presenting the conclusion covering the summary of the analysis.

Literature Review

Usage-feature analysis in Cognitive Semantics (Glynn, 2010b) attempts to make generalizations about the usage of words or lexemes, then make an assumption that certain usage patterns represent the speaker's knowledge of symbolic relations that denote "meaning". The usage-feature analysis is developed based on usage-based theory (Bybee, 2013). The principle of this theory is that cognitive representation of language is impacted and created by experience with language (Langacker 1987, 2000b; Kemmer & Barlow, 2000 as cited in Bybee, 2013). This theory highlights an essential characteristic of human language: high repetition of individual units and sequences of the unit leads to conventionalization, association, and automation of sequences (Bybee, 2013). Therefore, the frequency of co-occurrences between words from usage data distinguishes the meaning of words (Rajeg et al., 2020).

Despite the claim that this methodology is more empirical than the traditional approach towards meaning, several critics arise concerning corpus representation, quantification of meaning, and reliance on frequency to understand language. With respect to corpus representation, it is argued that enumeration as a goal can never be achieved due to the infinity of language (McEnery & Wilson, 2005). McEnery and Wilson (2005) further elaborated on the critics of the methodology that since language examples in natural language is infinite, corpora are considered incomplete in nature and cannot represent language. Based on this view, the empirical approach toward language is questioned for its reliability in representing natural language. Moreover, as the corpus is related to language usage or performance, the schools of Structuralist and Mentalist argued that an object of study should be the speaker's competence. Because of this, introspection is considered an ideal method (Glynn, 2010a). In relation to the measurement of meaning, this methodology is questioned regarding the effort of applying quantitative techniques (Glynn, 2010a). As meaning is an abstract concept of subjectivity and a non-observable phenomenon, it requires introspection. While introspection data are considered unreliable, Broekhuis (2020) claimed that most of the data are based on intersubjectivity, which means that the judgment is also built on other people's judgment when the researchers are not entirely sure about their own. Lastly, on the issue of frequency, some linguists argue that it can never be used to describe a language, especially regarding salience. Hence, the reliability of the usage-based approach is still debatable in the study of meaning (Glynn, 2010a).

While aware of these questions, Glynn (2010a) and Geeraerts (2010), however, argue that this methodology is reliable for several reasons. First, the usage-based approach is necessarily subjective because introspection is essential in corpus linguistics. Therefore, using a quantitative approach in the usage-feature analysis is not about *p-values*. It involves meticulous analysis of a wide range of formal, semantics, and sociolinguistic features of natural language that consist of thousands of examples. The manual or semi-automatic analysis leads to the availability of quantitative treatment of the data. Therefore, *p-values* are just one of the other important roles that indicate statistical significance. Second, quantification in semantics offers several benefits regarding confirmation of the statistical significance of the result, identification of patterns in usage that cannot be effectively identified by introspection, and accuracy test of an analysis (Glynn, 2010a). However, what may serve as the weakness of this approach is that since it requires a large number of instances, a small number of data may affect the result. It is because analysis of meaning requires a labor of intensive manual annotation, which may restrict the number of samples. With a small number of samples, the number of different factors being examined in the study will be restricted, thus affecting the result (Glynn, 2010b). Therefore, to have a more reliable result, a more significant number of data might be needed.

In order to make semantic data possible for quantification, a previous study of social emotion using the introspective method is used as a theoretical reference, in which hypotheses are formulated and conceptual roles that play a scenario of pain are developed. Albeit the absence of relevant study on the semantic aspect of pain in Indonesian, Goddard and Wierzbicka (2014) claimed that it is a human universal because people across the world feel it in some ways. Therefore, the cross-linguistics concept of pain is used here as the gate to go into the conceptual structure of pain in Indonesian. The attempt to test a hypothesis based on the introspective study of meaning is mentioned by Geeraerts (2010). He argued that subjective analyses can be used as the first step in the cycle of analysis because they represent hypotheses, therefore, they should be tested.

The present study provides a cross-cultural comparison of the study on the emotion of pain. Pain is considered culture-specific as a part of emotion (Goddard, 2011). However, Goddard and Ye (2016) argued that cultural diversity is often underestimated, so when the word is used without considering its cross-cultural difference, it may risk "a biased discourse that is centered on the Anglo cultural perspective" (p. 2). Due to its specified concept, it hardly has equivalence in other languages that do not have similar concepts to that of English. In order to address this issue, the natural semantic metalanguage (NSM) concept of "someone feels something bad" is used to explain the universality (Wierzbicka, 2012, p. 311). While the prototypical concept of pain in English covers bad feelings in one's part of the body due to a concurrent bad event that includes the person's consciousness and the unwanting feeling, such a similar concept may be absent in other languages. To illustrate this, Wierzbicka (2012) explains the absence of the concept of pain in Yankunytjatjara, one of the Australian Aboriginal languages. Owing to the absence of an exact counterpart of the word pain in English, the closest concept suggested by the dictionary of the language is the noun pika, which is related to "a bad feeling located in a particular part of the body and associated with physical cause" (Goddard & Wierzbicka, 2014,

p. 142). This concept does not construe the psychological pain expressed in English (Goddard & Wierzbicka, 2014; Wierzbicka, 2012).

Another cross-linguistic perspective of pain is conceptualized by a single word, *dolor* in Spanish. While in English, pain is described via pain-related words such as sore, ache, pain, and hurt, in Spanish, all localized bodily symptoms and emotional suffering can be expressed by the word *dolor* (Silva, 2016). This is in contrast with Indonesian, in which pain can be experienced via several pain-related lexemes such *as sakit, nyeri, pedih, perih, linu, and ngilu*. However, in this paper, the focus of discussion only covers the two lexemes: *sakit* and *nyeri*. Although pain in Indonesian is less explored, its corresponding meaning can be found in Tagalog and Cebuano-Bisayan's word for pain: *sakit*. As the word etymologically comes from Proto-Malayo-Polynesian (Wictionary), the comparison of the meaning of the word in different languages can be identified. In Tagalog, *sakit* can either be related to a localized pain in the body caused by a medical condition or refer to a broader phenomenon that involves well-being which refers to emotional pain (Cordero, 2021; de Castro & Alvarez, 2004). Similarly, *sakit* in Cebuano-Bisayan also covers pain on the part or whole of the body as expressed in the saying *Ang sakit sa kumingking sakit pud sa tibu "uk lawas*". "The pain experienced by a little finger is the pain as well of the whole body" (Lanaria, 2009, pp. 65-66). Moreover, the dictionary definition of *sakit* is "ache, emotional pain" in *Ang sakit sa ákung kasingkásing.* "The ache in my heart" (Wolff, 1972, p. 848), as well as a figurative expression of *Ang akung kasing-kasing* "My heart is pained" (Lanaria, 2009, p. 61) also clearly embodies the concept of emotional pain. The physical and non-physical concepts of bad feelings in the body likely indicate the similar meaning of *sakit* in both languages in which the discussion mainly focuses on two dimensions: bodily focus (physical or emotional) and location (part or whole of the body) with no further elaboration on intensity and duration.

As the conceptualization of pain varies across languages in which the diversity is viewed from a cultural perspective, the diverse concept can be explained by the Saphir-Whorf hypothesis on linguistic diversity. Linguistic relativity proposes that "the particular language that one speaks influences the way one thinks about reality" (Lucy, 2001, p. 903). This theory highlights two critical claims, which are viewed as strong and weak hypotheses. The strong hypothesis is related to linguistic determinism, which strongly argues that the way people think is determined by language. In contrast, the weak hypothesis is related to the partial influence of language on thought (Jiang, 2017). Concerning language determination on thought, Wierzbicka (1997) argued that a person's native language influences his or her conceptual perspective on life. The influence of language on the way one perceives reality is closely linked to the way the vocabulary of a particular language treats a specific phenomenon (O'Neill, 2006). Wierzbicka (1997) supported this argument by providing evidence on how the concept of freedom in Roman (libertas) and Russian (svoboda) is different because of culture and history. While libertas presents a concept of one as a master on his/her own and not under someone else's control, svoboda suggests no external constraints on someone's action and a sense of well-being. Therefore, the different conceptualization of the meaning of words between languages illustrates the strong influence of one's native language on their thinking habits (Wierzbicka, 1997).

Referring to the cross-linguistic perspective of pain, Wierzbicka (2016) proposes several concepts as the basic idea of pain: "someone can feel something bad in their body." Although the concept of pain varies cross-culturally, the differences may include the following dimensions. First, pain can either focus on the body or the person as a whole. Second, a reference to the body as the locus of pain can refer to the body as a whole or as a part of the body. Third, regarding intensity, the bad feeling can refer to someone who feels something bad or something very bad in the body. Lastly, pain can vary by "bad feeling" of any duration, and bad feeling extended in time. This study uses these dimensions to code each concordance of the lexeme from corpus data before revealing their conceptual differences using quantitative analysis.

Based on the theoretical framework of usage-based theory (Bybee, 2013) and the cross-linguistic concept of pain (Wierzbicka, 2016), this study is contextualized by the hypothesis on the semasiological structure of pain from the Indonesian dictionary. If referring to Online Indonesian Dictionary (KBBI) (Badan Bahasa, n.d.), *sakit* is related to pain in the body or part of the body because of fever, stomachache, and so forth. At the same time, *nyeri* is also related to pain in the body. However, the dictionary describes it as the feeling when part of the body is injected repeatedly or clamped, which causes suffering (Badan Bahasa, n.d.). Moreover, KBBI also lists the second sense of *nyeri* from the psychology domain, defining *nyeri* as a physical and emotional experience that is caused by tissue injury (Badan Bahasa, n.d.). The definition of these two lexemes from the dictionary likely indicates that the two lexemes of pain can be differentiated by the focus of pain, degree or intensity, as well as duration. By degree of pain, *sakit* is likely to be less painful, or it can be assumed that *nyeri* is more painful and has a more lasting effect. Therefore, this study hypothesized that *sakit* and *nyeri* are distinguishable by location, intensity, and duration dimensions.

.

Methods

This study employs a corpus-driven method by using usage-feature analysis (Glynn, 2010b). Two statistical analyses were conducted namely exploratory and confirmatory statistics. In relation to exploratory analysis, Binary Correspondence Analysis (Glynn, 2014) was used to find the co-occurrence of usage features that gives a map of the patterning of the two lexemes (Glynn, 2014). Glynn (2014) noted that Correspondence Analysis does not require equal distribution, but making a balanced selection for each form is the best way of achieving it. Meanwhile, to confirm the significance of the patterns, Binomial Regression Analysis (Levshina, 2015) was employed. The statistical analysis was conducted in R (R Core Team, 2022) using the code provided by Glynn (2014) and Levshina (2015).

Data Collection

Corpus data used in analysis and presented as instances in the discussion was downloaded from the Leipzig Corpora Data Collection of Indonesian (Goldhahn et al., 2012). The instances from Leipzig Corpora Data Collection of Indonesian News 2020 that are included in this paper are provided with their English translations. Leipzig Corpora is a web-crawled corpus. The data are sourced from news publishers, newspaper collections, web-based news, Wikipedia, and text retrieved from websites (Biemann et al., 2007). The genre of news has been mainly used in this analysis. Approximately 200 examples were randomly extracted for each lexeme, with a total of 400 sentences. The rationale behind the number of examples solely relies on the feasibility of the analysis. Compared to similar studies that have been conducted, Krawczak (2015) used approximately 400 instances, while (Glynn, 2010b) used in total 650 instances. It shows that the number of instances values the quality of the result.

Data Analysis

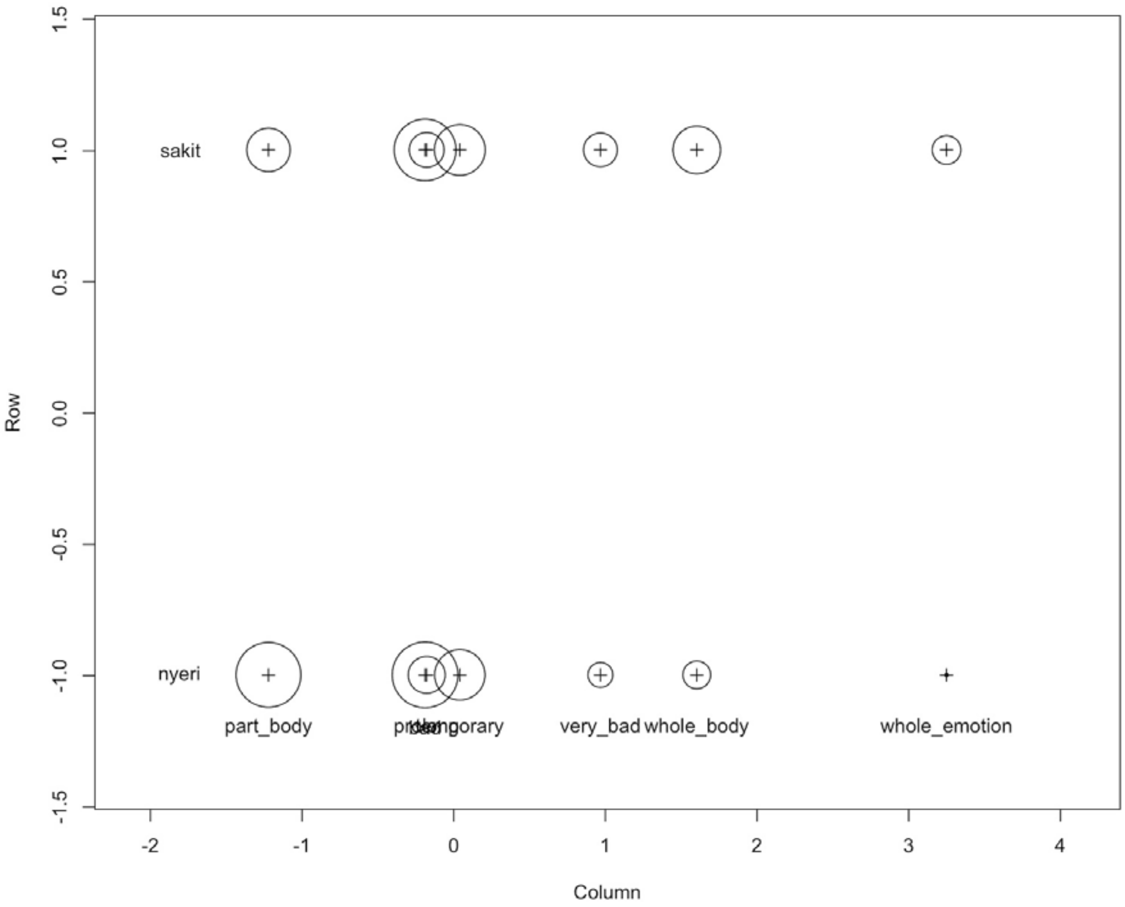
Analysis was conducted based on the following steps. First, the corpus data of Indonesian News 2020 was downloaded from the Leipzig Corpora Data Collection of Indonesian (Goldhahn et al., 2012) with a total of 15,206,710 tokens. Then, the corpus data was processed in AntConc version 4.1.2 (Laurence, 2022) for searching the concordance list by using an advanced setting. The focus of the search was limited to the context of verbal and adverbial expressions related to "feeling" such as *rasa, merasakan, derita, menderita, makin, semakin,* and *mulai* in order to avoid hits of *rumah sakit* "hospital", which is not directly related to *sakit* as a bad feeling. A total of 400 examples were extracted from AntConc (Laurence, 2022) before manual annotation. The data were coded manually based on dimensions of pain, such as bodily focus (part or whole), intensity (bad or very bad), and duration (momentary or prolonged). After annotation, the data was loaded into R (R Core Team, 2022) for Binary Correspondence Analysis using R package *{MASS}* (Glynn, 2014). After that, the analysis proceeded in Binomial Logistic Regression (Levshina, 2015) using *lrm* function. As stated by Speelman (2014), this confirmatory analysis can be used to identify factors that have an impact on the choice between near-synonyms as well as separating their respective effects.

Results

This section covers the result of two statistical analyses. The result of the exploratory analysis (Glynn, 2014) is presented in a graphic representation showing the clustering of usage feature of the two lexemes. The next result of confirmatory analysis from Binomial Logistic Regression shows the reliability and accuracy of the profiles to approve the identified conceptual association of the lexemes that have been annotated manually.

The result of the Correspondence Analysis is presented in Figure 1. The result is plotted in a biplot with *nyeri* and *sakit* on the x-axis (representing the row), while the dimensions of pain are plotted on the y-axis (representing the column). The clustering of the dimension of pain consists of three levels of bodily focus (*part_body*, *whole_body*, and *whole_emotion*), two levels of intensity (*bad* and *very_bad*), and two levels of duration (*temporary* and *prolonged*). Bodily focus represents the physical and non-physical dimensions of pain, while intensity shows the degree of pain, whether it is bad or very bad. Lastly, duration describes the usage feature related to the length of pain, whether it is momentarily or prolonged. The biplot presents the categories of the variables in the positive and negative values on both axes, indicating the association between them. On the x-axis, *sakit* is plotted on a positive value while *nyeri* is on a negative value. Meanwhile, the categories of dimension are plotted along the y-axis, where *part_body* is clearly plotted on the negative value, while *very_bad*, *whole_body* and *whole emotion* are on the positive value. The overlapping categories of *prolong*, *bad*, *temporary* are plotted nearly at the center of the plot.

Figure 1
The Result of Correspondence Analysis



Based on the plot, it appears that *nyeri* is more strongly associated with *part_body* than with *whole_body* and *whole_emotion*. The distance between *nyeri* and *part_body* is shorter than the distance between *nyeri* and *whole_body*, indicating a stronger association with the former. Meanwhile, *sakit* is located closer to *whole_body* and *whole_emotion* compared to the other categories. This suggests a stronger association between *sakit* and these two categories compared to the other categories. Moreover, *prolong*, *temporary*, and *bad* are located near the center of the plot, indicating that they are not strongly associated with any of the categories in the analysis. However, based on the distance, *temporary* is likely closer to *whole_body* and *whole_emotion* categories compared to other categories. This may suggest that pain that is described as *temporary* is more likely to involve the entire body and emotions, rather than being focused on a specific part of the body or limited to a physical sensation.

The distance of the categories towards certain lexeme may indicate that *sakit* and *nyeri* can be differentiated by bodily focus and intensity. It can be assumed that *nyeri* is closely related to pain in a specific part of the body, while *sakit* is closely related to pain in the whole body. The following examples from the corpus data illustrate the two distinct usage features. The association of *sakit* with pain in the body or person as a whole (coded as whole_body) is presented from (1) to (6), while *nyeri* with pain in the part of the body is explicated from (7) to (12).

1. Gangguan somatik merupakan salah satu gangguan mental di mana sang penderita mengeluh rasa sakit di tubuhnya namun tidak dapat diketahui penyebabnya.
   Somatic disorder is a mental disorder in which the sufferer complains of pain in his body but cannot identify the cause.

2. Dia telah diberikan obat pembunuh rasa sakit agar bisa meredakan rasa sakit yang dialaminya.
   S/he had been given pain killers to relieve the pain s/he was experiencing.

3. Tidak biasa bagi perempuan untuk tidak meminta epidural untuk menghilangkan rasa sakit selama persalinan, terutama ketika melahirkan untuk pertama kalinya.
   It's uncommon for women not to request an epidural for pain relief during labour, especially when giving birth for the first time.

4. Dia mengaku sempat menahan rasa sakit sejak bulan lalu hingga akhirnya memeriksakan diri di rumah sakit.
   S/he admitted that s/he had endured the pain since last month until finally having a medical checkup in the hospital.

5. Gara-gara kebiasaan buruk itulah kini Betrand harus menahan sedikit rasa sakit pada perutnya.
   Because of his bad habit, now Betrand has to endure a little pain in his stomach.

6. Dia mengalami sedikit rasa sakit di bahunya.
   S/he had a slight pain in his/her shoulder.

7. Selain itu motif lainnya, menurut Ramadhan, ada yang didorong karena rasa sakit hati dan dendam sehingga melalukan tindakan pengeroyokan dan penganiayaan, bahkan sampai pembunuhan.
   In addition to other motives, according to Ramadhan, some were driven out of hurt (resentful or lit. heart pain) and revenge so they carried out acts of beatings and persecutions, even to the point of murder.

8. Dia mengatakan ada rasa sakit yang mendalam terukir di jiwa bangsa kita dan di hati jutaan orang.
   He said there is a pain that is deeply engraved in the soul of our nation and in the hearts of millions.

9. Stress dapat memicu produksi asam lambung meningkat, sehingga dinding lambung akan mengalami iritasi, maka timbul rasa nyeri yang berlebihan di lambung.
   Stress can trigger stomach acid production, so that stomach wall will experience irritation, resulting in excessive pain in the stomach.

10. Gejala gigitan semut api biasanya diawali dengan rasa nyeri yang sangat tajam, seperti terbakar atau habis dicubit.
    Symptoms of a fire ant bite usually begin with a very sharp pain, like burning or being pinched.

11. Tak heran dia merasakan nyeri yang berlebihan di bagian perut.
    No wonder he felt excessive pain in the stomach.

12. Operasi itu seharusnya dapat meredakan nyeri punggung kronis.
    The surgery was supposed to relieve chronic back pain.

13. Hal ini mengakitbatkan kristal menumpuk di persendian dan menyebabkan sendi terasa nyeri, bengkak, dan meradang.
    This causes crystals to build up in the joints and causes the joints to feel painful, swollen, and inflamed.

14. Olahraga ini dapat mengurangi nyeri haid dan ketidaknyamanan selama menstruasi.
    These exercises can reduce menstrual pain and discomfort during menstruation.

The usage feature of *sakit* shows that *sakit* can be associated with pain in the part of the body, such as in (5) and (6). Yet, the exploratory analysis shows the tendency of pain experienced by a person as a whole, which is illustrated in (1) to (4). In comparison with *sakit*, the whole bodily sensation rarely occurs with *nyeri*. This is best explicated in (9) to (14), in which *nyeri* is associated with pain in the stomach or lower abdomen (9, 11, 14), part of the body which is bitten by a fire ant (10), the back of the body (12), and the joints (13). Meanwhile, emotional pain only corresponds to *sakit* with no occurrences in *nyeri*. This is best illustrated by (7) and (8), in which *sakit* is related to emotional sensation in the heart and soul. In terms of intensity, although the very bad intensity is plotted relatively close to *sakit*, the dimension can be identified in both *sakit* and *nyeri.* The very bad intensity in *sakit* is explicated in (2) and (3) for physical pain and in (8) for the emotional phenomenon. Comparable with *sakit*, the intensity in *nyeri* can be identified by the use of intensifiers, such as *yang berlebihan* “excessive” (9) and *yang sangat tajam* “very sharp” (10). Moreover, due to its very bad intensity, it needs to be relieved, as explicated in (12) and (14). While the two dimensions of pain show a tendency to a certain lexeme, the duration of pain shows no distinct usage between the two lexemes.

The exploratory analysis has given a particular usage pattern to specific lexemes. However, the degree of accuracy needs to be confirmed. The accuracy and predictive power of the findings are confirmed by Binomial Logistics Regression. The result is presented in Table 1.

Table 1
The Result of Binomial Logistic Regression (Model 1)

lrm (formula = node ~ body_focus + intensity + duration, data = dat)				
	Model Likelihood Ratio Test	**Discrimination Indexes**	**Rank Discrim. Indexes**	
Obs 309	LR chi2 95.02	R2 0.353	C 0.766	
*nyer*i 156	d.f. 4	R2(4,309) 0.255	Dxy 0.532	
sakit 153	Pr(>chi2)<0.0001	R2(4,231.7) 0.325	Gamma 0.673	
max |deriv|0.001		Brier 0.184	tau-a 0.267	
	Coef	**S.E.**	**Wald Z**	**Pr(>|Z|)**
Intercept	-1.3393	0.3049	-4.39	<0.0001
body_focus=whole_body	1.9815	0.3142	6.31	<0.0001
body_focus=whole_emotion	11.0543	30.5787	0.36	0.7177
intensity=very_bad	0.5666	0.4057	1.40	0.1625
duration=temporary	0.7129	0.3204	2.23	0.0261

The result of analysis using *lrm* function on R (R Core Team, 2022) gives a number of statistical values related to the accuracy of the model as well as the predictors. The column of Model Likelihood Ratio Test shows that the model is significant in general with p-value <0.0001. The value on statistics C=0.766 means that the model has acceptable discrimination (Levshina, 2015). However, the model shows that there is an insignificant predictor (intensity=very_bad) with a p-value of 0.16. In order to get accuracy related to significant predictors, the model is then pruned manually in a step-wise fashion.

The final model (Model 2) is presented in Table 2, showing the whole body as the only significant predictor with a p-value <0.0001. The model sets part of the body by default as the referent level with estimated log odds -0.79 or simple odds 0.45, meaning that the chances of *sakit* are 0.45 times smaller than *nyeri* in the context of part of the body. Whereas the coefficient of significant predictor is in a positive value, meaning that the usage feature of the whole body favors *sakit* in comparison to part of the body. The log odds ratio of the significant predictor is 1.88 (simple odds=6.53). This value means that the chances of *sakit* compared to *nyeri* in the context of the whole body are 6.53 times higher than those in the context of part of the body.

Table 2
The Result of Binomial Logistic Regression (Model 2)

lrm(formula = node ~ body_focus, data = dat)				
	Model Likelihood Ratio Test	**Discrimination Indexes**	**Rank Discrim. Indexes**	
Obs 400	LR chi2 112.78	R2 0.328	C 0.745	
nyeri 200	d.f. 2	R2(2,400)0.242	Dxy 0.491	
sakit 200	Pr(>chi2) <0.0001	R2(2,300)0.309	gamma 0.806	
max |deriv| 0.001		Brier 0.189	tau-a 0.246	
	Coef	**S.E.**	**Wald Z**	**Pr(>|Z|)**
Intercept	-0.7992	0.1381	-5.79	<0.0001
body_focus=whole_body	1.8760	0.2496	7.51	<0.0001
body_focus=whole_emotion	11.0020	28.5945	0.38	0.7004

Discussion

In general, it can be assumed that logistic regression confirms the exploratory analysis to a certain degree. First, the usage feature of the two near-synonyms is distinguished by body focus, in which *sakit* is more associated with pain in the whole body while *nyeri* is related to pain in a specific part of the body. Despite the association with the body in general, *sakit* is also a localized bodily sensation. This allows expression such as *sakit* kepala "headache" and *sakit* gigi "toothache" in which pain is expressed by locating it to a particular part of the body with reference to illness or diseases. The association of pain as a localized bodily sensation is more comparable with Tagalog (Cordero, 2021; de Castro & Alvarez, 2004), in which *sakit* is always located based on where the pain is felt. The association of *sakit* with both a localized and a general bodily sensation is comparable to the usage in Cebuano-Bisayan (Lanaria, 2009; Wolff, 1972). In contrast, *nyeri* is more associated with a localized feeling, so the localized bodily expressions such as *nyeri punggung* "back pain" and *nyeri sendi* "joint pain" are more common than *rasa nyeri di badan* "pain in the body". When referring to pain in the body as a whole, *sakit* is used instead of *nyeri*, so it is more common to say *Badan saya terasa sakit* "My body is painful" than *Badan saya terasa nyeri** "My body is painful". This general conceptualization of *sakit* compared to *nyeri* is more associated with pain in English. While pain in English is related to something bad that is happening to someone's part of the body, Goddard and Wierzbicka (2014) argued that sometimes it could be conceived as global. Therefore, an expression such as *She is in pain*, in which pain does not refer to any particular part of the body, is comparable to the Indonesian expression "Dia kesakitan" with no comparable expression in *nyeri* such as "Dia kenyerian"*.

Moreover, although emotion is not a significant predictor in the model, the exploratory analysis shows that the usage feature is mostly related to *sakit* without any chances of occurrences in *nyeri*. It can be assumed that *sakit* is related to physical and emotional pain, while *nyeri* is only related to physical pain. This finding may also correspond to the occurrence of the idiomatic expression *sakit hati* (lit. ill liver) in Indonesian that is related to displeased feeling such as revenge and hatred due to being hurt emotionally. The association of *hati* in Indonesian and

Malay as the center of emotion and thought is explained by Siahaan (2008) and Goddard (2008; 2010). Siahaan (2008) argued that the anatomical use of liver as the conceptualization of emotion compared to heart in English has been influenced by ancient Indonesian ritual of liver divination and ancient cultural belief. Meanwhile, Goddard (2008; 2010) who analysed *hati* in Malay emphasized it as the key word of Malay culture, particularly as the conceptualization of emotion. Because of this, the fixed expression of *sakit hati* is related to "emotional hurt rather than the illness in this organ" (Siahaan, 2008, p. 55). Similarly, Goddard (2010) also explained that this expression may imply something like pain. While pain in the heart or feeling is closely related to *sakit* with no similar expression that corresponds to *nyeri*, it is possible that the usage feature of emotion does not occur in *nyeri*. However, the definition in the dictionary shows that emotional pain in *nyeri* is related to the psychological domain. The absence of the usage feature of emotion in *nyeri* may also be related to the genre of the corpus in this study. Since the genre is only news, it may also decrease the possibility of occurrences in relation to emotional pain in *nyeri* due to psychological effects because such domain may be related to specific genres such as academic journals.

In relation to the hypothesis, the finding confirms but also disconfirms the hypothesis in this study. The result of statistical analysis confirms the hypothesis that *sakit* and *nyeri* can be distinguished by the usage feature of body focus. While sakit can occur in the body as a whole and the part of the body, the association is statistically related to pain in the whole body compared to the part of the body, which is more associated with nyeri. However, it disconfirms that *nyeri* is more painful and lasting than *sakit* because intensity and duration are not significant in the final model of confirmatory analysis.

Overall, this result gives a discriminatory picture of the use of near-synonyms of pain for L2 learners of Indonesian. The difference in meaning between the two lexemes can be clearly identified by the context of body focus. Based on bodily focus, *sakit* is related to pain in the whole body, which covers part of the body as well. In other words, *sakit* is a word used to describe the unpleasant bodily feeling in general. This context of use allows an expression like *Dia menunjuk kaki Luna yang sakit dan masih diperban* "He pointed Luna's leg which was painful and was still bandaged", in which pain is localized. Moreover, *sakit* can be used for unpleasant bodily sensation in general, such as *Lewat pengalaman, pertemuan rasa sakit dan bahagia, kita dituntun untuk menjadi insan yang lebih baik* "Through experience, the crossing of pain and happiness, we are led to become a better person". From this example, *sakit* can be related to either non-specific unpleasant bodily sensation or unpleasant non-physical sensation, which corresponds to emotion. It enables the use of *sakit* in a more general context, not only bodily sensation. Meanwhile, although temporary duration shows no distinct usage between the lexemes, it is likely related to *sakit* relative to *nyeri* if compared to *hurt* in English. Wierzbicka (2014) mentioned that *hurt* and *pain* do not essentially mean the same in English. While *pain* is global, *hurt* is a localized bodily and short-time occurrence. However, in Indonesian, *sakit* may cover the meaning of both *pain* and *hurt* as it can refer to a generally unpleasant bodily sensation in the previous example as well as a localized bad feeling on body that occurs temporarily after an immediate preceding cause as in the sentence like *Ditusuk jarum yang harusnya terasa sakit, justru dirasakan Westny hanya terasa dingin* "Being stabbed by a needle, which should have hurt, Westny only felt cold".

Meanwhile, *nyeri* is more localized and physical as it occurs most commonly on specific parts of the body, such as muscles, chest, epigastrium, and joints. The cause of this physical pain may cover chronic disease in a sentence like *Biasanya penderita asam urat juga merasakan nyeri di bagian-bagian tertentu seperti jari tangan, lutut, pergelangan tangan* "Usually gout sufferers also feel pain in specific parts such as fingers, knees, wrists". It is also experienced because of injury, which is illustrated in a sentence like *Korban mengalami luka cakaran di lengan dan wajah, serta nyeri di beberapa bagian tubuh* "The victim suffered claw wounds on his arms and face, as well as pain in several parts of his body". Moreover, although the duration is not a significant predictor based on statistical analysis, in some contexts, *nyeri* likely appears to be more lasting than *sakit,* as expressed in the following sentence *Beberapa pasien dilaporkan hanya kelelahan, menggigil, sakit kepala, dan nyeri di tempat suntikan* "Some patients were reported only felt fatigue, chill, headache, and pain at the injection spot". It may suggest that *nyeri* has a more lasting effect compared to *sakit* in relation to injection. While *sakit* is a short-period experience after an immediate occurrence such as an injection, *nyeri* occurs longer as the side effect of the action.

Implication

The result of this study gives a new insight for L2 learners of Indonesian on the study of near-synonyms from a cross-cultural perspective. Since the meanings of emotion words are considered to be culture-specific (Goddard and Ye, 2016), as an implication, this study may enrich L2 vocabulary learning and cross-cultural competence. This study demonstrates the use of corpora to study near-synonyms of pain in Indonesian as well as comparing them with other languages. The finding, thus, enriches L2 lndonesian learning for both of educators and learners. For educator, this study offers possibility of combining cross-cultural competence and vocabulary learning in L2 learning design by making use of corpora. Moreover, for L2 Indonesian learners, this study may have an immediate implication in vocabulary learning as the finding helps them discriminate near-synonyms of pain based on the context of use as well as presenting the comparison of meaning with other languages.

Recommendation

This study has demonstrated using a corpus-driven approach to study near-synonyms in Indonesian. As this study may benefit L2 learners of Indonesian, several recommendations are made for practical purposes and methodology development for future study. For the practitioner, a similar approach can be used to test the usage features of near-synonyms in other languages. As many corpora have been made available for public use, it may provide a rich context of use to help L2 students differentiate near-synonyms based on their usage. In terms of methodological scope, despite giving a clear context on the use of near-synonyms, the present study can be developed for further analysis. First, concerning the genre of the corpus, this study illustrates the use of a specific genre of corpus to analyze emotion. Future studies on emotion using a similar methodology can be developed by comparing usage profile of lexemes from different genres of the same corpus or different corpora, for example, between BNC and Kolhapur. Specifically, for the study of pain in Indonesian, the current study can be extended by using specific genres, such as the psychological or medical domain. Since corpus data in this study only covers the genre of news, expanding the corpus composition may enrich the finding of this study. Moreover, in relation to variables, a similar study can be adapted to more pain-related lexemes in Indonesian. The additional variables may enrich the cross-linguistic concept of pain and help L2 learners of Indonesian understand the context of use to achieve the success of cross-cultural communicative competence.

Conclusion

Sakit and *nyeri* are two adjectives of pain that have closely related meaning in Indonesian. By using the multivariate corpus-driven approach, the analysis shows that *sakit* is more associated with pain in the whole body and emotions with very bad intensity. At the same time, *nyeri* is more associated with pain on the part of the body. However, the confirmatory analysis indicates that body focus is the only significant predictor of the near-synonyms. This finding confirms the hypothesis that *sakit* and *nyeri* can be distinguished by bodily focus. While *sakit*, by dictionary definition, covers both part and whole body-focus, it is likely more associated with the whole body than *nyeri*, which only covers specific parts of the body. The general focus of *sakit* includes physical and emotional pain, while the specificity on *nyeri* only covers physical phenomena.

Moreover, the confirmatory analysis disconfirms the hypothesis that *nyeri* is more painful and lasting than *sakit*. Despite the result, the analysis of this study has limitations. First, concerning corpus size, the data in this study is relatively small compared to other similar studies (Glynn, 2010b; Krawczak, 2015; Rajeg, 2016; Rajeg et al., 2020). Therefore, it may not cover all factors proportionally, particularly occurrences of intensity and duration. Secondly, related to a specific genre, this study only uses the news genre. Thus, it may decrease the possibility of occurrences of emotional pain in *nyeri* related to the psychological domain, which is more scientific. Therefore, future study with more significant instances and more varied genres of the corpus is recommended for a more reliable result.

Acknowledgement

I want to thank Dr Ksenia Gnevsheva for her invaluable comment on the draft of this paper during the Quantitative Research Method Class at the Australian National University. I thank the Indonesian Endowment Fund for Education (LPDP) from the Ministry of Finance, the Republic of Indonesia, for granting the scholarship. All the remaining deficiencies are my sole responsibilities.

References

Badan Bahasa. (n.d.). *Kamus Besar Bahasa Indonesia.* Retrieved 31 August 2022, from https://kbbi.kemdikbud.go.id

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, *2007.*

Broekhuis, H. (2020). Why I will not become a corpus linguist: The use of introspection data and corpus data in synchronic syntactic research. *Nederlandse Taalkunde*, *25*(2–3), 181–192. https://doi.org/https://doi.org/10.5117/NEDTAA2020.2-3.003.BROE

Bybee, J. L. (2013). Usage-based Theory and Exemplar Representations of Constructions. In *The Oxford Handbook of Construction Grammar*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0004

Byram, M., & Wagner, M. (2018). Making a difference: Language teaching for intercultural and international dialogue. *Foreign Language Annals*, *51*(1), 140–151. https://doi.org/https://doi.org/10.1111/flan.12319

Cordero, D. A. (2021). Sákit pighati and pag-asa: A pastoral reflection on suffering during the COVID-19 pandemic in the Philippines. *Journal of Religion and Health*, *60*(3), 1521–1542. https://doi.org/10.1007/s10943-021-01234-5

de Castro, L. D., & Alvarez, A. A. A. (2004). Sakit and Karamdaman: Towards authenticity in Filipino concepts of disease and illness. *Bioethics: Asian Perspectives: A Quest for Moral Diversity*, 105–112. https://doi.org/10.1007/978-94-017-0419-9_10

Dorsey, B. F., Kamimura, A., Cook, L. J., Kadish, H. A., Cook, H. K., Kang, A., Nguyen, J. B. T., & Holsti, M. (2022). Communication gaps between providers and caregivers of patients in a pediatric emergency department. *Journal of Patient Experience*, *9*, 237437352211122-23743735221112223. https://doi.org/10.1177/23743735221112223

Geeraerts, D. (2010). The doctor and the semantician. In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* De Gruyter, Inc. https://doi.org/10.1515/9783110226423.61

Glynn, D. (2010a). Corpus-driven Cognitive Semantics Introduction to the field In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* De Gruyter, Inc. https://doi.org/10.1515/9783110226423

Glynn, D. (2010b). Testing the hypothesis: Objectivity and verification in usage-based Cognitive Semantics. In *Quantitative Methods in Cognitive Semantics. Corpus-Driven Approaches*. Mouton de Gruyter. https://doi.org/10.1515/9783110226423.239

Glynn, D. (2014). Correspondence analysis: Exploring data and identifying patterns. In Glynn, D. & Robinson, J. (Eds.), *Quantitative Studies in Polysemy and Synonymy*. (pp. 443–486). https://doi.org/10.1075/hcp.43.17gly

Goddard, C. (2008). Contrastive semantics and cultural psychology: English heart vs. Malay hati. *Sharifian, Farzad, Dirven, Rene, Ning Yu, & Niemeier, Susan (Eds). Culture, Body, and Language: Conceptualizations of Body Organs across Cultures and Languages. The Hague: Mouton de Gruyter*, 75–102.

Goddard, C. (2010). Hati: A key word in the Malay vocabulary of emotion. In *Emotions in Crosslinguistic Perspective* (pp. 167–196). De Gruyter Mouton. https://doi.org/10.1515/9783110880168.167

Goddard, C. (2011). *Semantic analysis: A practical introduction*. Oxford University Press.

Goddard, C., & Wierzbicka, A. (2014). *Words and meanings lexical semantics across domains, languages, and cultures* (In First Edition ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199668434.001.0001

Goddard, C., & Ye, Z. (2016). Exploring "happiness" and "pain" across languages and cultures In C. Goddard & Z. Ye (Eds.), *"Happiness" and "Pain" across Languages and Cultures* (pp. 1–18). John Benjamins Publishing Company. https://doi.org/10.1075/bct.84.01god

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Liepzig Corpora Collection: From 100 to 200 Languages *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*.

Jiang, S. (2017). Linguistic relativity and empirical studies. In (1st Ed., pp. 29-42). Routledge. https://doi.org/10.4324/9781315265483-3

Kalengayi, F. K. N., Hurtig, A.-K., Ahlm, C., & Ahlberg, B. M. (2012). "It is a challenge to do it the right way": an interpretive description of caregivers' experiences in caring for migrant patients in Northern Sweden. *BMC health services research*, *12*(1), 1–18. https://doi.org/10.1186/1472-6963-12-433

Krawczak, K. (2015). Shame and its Near-synonyms in English : A Multivariate Corpus-driven Approach to Social Emotions.

Lanaria, L. L. (2009). Lawas: An Anthropo-theological Discourse on the Body in A Cebuano-Visayan Language Context. *Philippine quarterly of culture and society*, *37*(1), 55–82.

Laurence, A. (2022). AntConc (Version 4.1.2) [Computer Software]. In. Tokyo, Japan: Waseda University.

Levshina, N. (2015). *How to do linguistics with R data exploration and statistical analysis*. John Benjamins Publishing Company,. https://doi.org/10.1075/z.195

Lucy, J. A. (2015). Sapir-Whorf Hypothesis. In (Second Edition ed., Vol. 20, pp. 903-906). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-097086-8.52017-0

McEnery, T., & Wilson, A. (2005). *Corpus Linguistics An Introduction* (Second Ed.). Edinburgh University Press.

O'Neill, S. (2006). Mythic and Poetic Dimensions of Speech in Northwestern California: From Cultural Vocabulary to Linguistic Relativity. *Anthropological linguistics*, *48*(4), 305–334.

Priestley, C. (2016). The semantics and morphosyntax of *tare* "hurt/pain" in Koromu (PNG) Verbal and nominal constructions In C. Goddard & Z. Ye (Eds.), *"Happiness" and "Pain" across Languages and Cultures* John Benjamins Publishing Company. https://doi.org/10.1075/bct.84.07pri

R Core Team. (2022). R: A language and environment for statistical computing In R Foundation for Statistical Computing (Ed.). Vienna, Austria.

Rajeg, G. P. W. (2016). Exploring the semantics of near-synonyms via metaphorical profiles: A quantitative corpus-based study of Indonesian words for HAPPINESS.

Rajeg, G. P. W., Rajeg, I. M., & Arka, I. W. (2020). Contrasting the semantics of Indonesian-kan &-i verb pairs: A usage-based, constructional approach. Seminar Nasional Bahasa Ibu.

Reppen, R., & Simpson-Vlach, R. (2020). Corpus Linguistics In N. Schmitt & M. P. H. Rodgers (Eds.), *Introduction to Applied Linguistics* (Third ed., pp. 91-108). Routledge Taylor & Francis Group. https://doi.org/10.4324/9780429424465-6

Siahaan, P. (2008). Did he break your heart or your liver? A contrastive study on metaphorical concepts from the source domain ORGAN in English and in Indonesian. *Culture, body, and language: Conceptualizations of internal body organs across cultures and languages*, *7*, 45–4.

Silva, Z. B. (2016). Some remarks on "pain" in Latin American Spanish. In C. Goddard & Z. Ye (Eds.), *"Happiness" and "Pain" across Languages and Cultures* (pp. 109–121). John Benjamins Publishing Company. https://doi.org/10.1075/bct.84.06bul

Speelman, D. (2014). Logistic regression A confirmatory technique for comparisons in corpus linguistics In D. Glynn, & J. A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy.* John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.18spe

Stone, M. (2023). Intersecting language and culture in the FL classroom. *Journal of Language Teaching and Research*, *14*(1), 1–5. https://doi.org/https://doi.org/10.17507/jltr.1401.01

Unger, S., Orr, Z., Avraham Alpert, E., Davidovitch, N., & Shoham-Vardi, I. (2022). Social and structural determinants and their associations with patient experience in the emergency department. *International emergency nursing*, *61*, 101131. https://doi.org/10.1016/j.ienj.2021.101131

Wictionary. (18 March 2023 03:28 UTC). *sakit*. Wictionary. Retrieved 21 April 2023, 17:06 from https://en.wiktionary.org/w/index.php?title=sakit&oldid=72160088

Wierzbicka, A. (1997). *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press, Incorporated.

Wierzbicka, A. (2012). Is pain a human universal? A cross-linguistic and cross-cultural perspective on pain. *Emotion Review*, *4*(3), 307–317. https://doi.org/10.1177/1754073912439761

Wierzbicka, A. (2016). "Pain" and "suffering" in cross-linguistic perspective. In C. Goddard & Z. Ye (Eds.), *"Happiness" and "Pain" across Languages and Cultures* John Benjamins Publishing Company. https://doi.org/10.1075/bct.84.02wie

Wolff, J. U. (1972). A dictionary of Cebuano Visayan. Volume One.

Yoon, K. J. (2007). My experience of living in a different culture: The life of a Korean migrant in Australia. In M. Besemeres & A. Wierzbicka (Eds.), *Translating Lives: Living with Two Languages and Cultures* (pp. 114–127). University of Queensland Press.

Yu, L.-C., Lee, L.-H., Yeh, J.-F., Shih, H.-M., & Lai, Y.-L. (2016). Near-synonym substitution using a discriminative vector space model. *Knowledge-Based Systems*, *106*, 74–84. https://doi.org/10.1016/j.knosys.2016.05.025

Corresponding author: Haniva Yunita Leo
Email: HanivaYunita.Leo@anu.edu.au