Constructing an Academic Thai Plagiarism Corpus for Benchmarking Plagiarism Detection Systems

Supawat Taerungruang
this.supawat@gmail.com
Chulalongkorn University, Thailand

Wirote Aroonmanakun
awirote@chula.ac.th
Chulalongkorn University, Thailand

ABSTRACT

Plagiarism is a major problem in the academic world. It does not only undermine the credibility of educational institutions, but also interrupts the processes of creating knowledge in the academic community. To lessen this problem, many plagiarism detection systems have been developed to detect plagiarized texts in academic works. In this paper, we describe the design and process in creating an academic Thai plagiarism corpus. This corpus is necessary for training and testing plagiarism detection systems for Thai. In order to make this corpus a comprehensive representation of plagiarism, the data has been divided into various types based on the degree of the linguistic mechanisms used in plagiarism. Data compiled in our corpus comes through two main methods: manually created by participants and automatically generated by a program. After the corpus is created, its validity is verified by using three measurements: a measurement of similarity between suspicious texts at the character level, a measurement of similarity between suspicious texts at the word level, and a comparison of different types of data compiled in the corpus based on the similarity measured. The results of the analyses indicate that the corpus created by the proposed methods is effective in training and testing plagiarism detection systems.

Keywords: plagiarism; Thai plagiarism detection; corpus creation; language resources; natural language processing

INTRODUCTION

Plagiarism is a major problem in the academic world. Not only does plagiarism undermine the credibility of educational institutions, but it also interrupts the processes of creating knowledge in the academic community (Chulalongkorn University, 2012, p. 3). Nowadays, information technology has been developed to be more effective. This makes it easy to access academic information and to take advantage of other people's ideas effortlessly. For this reason, the problem of plagiarism is intensifying as well (Sutherland-Smith, 2008, p. 75).

To handle plagiarism problems, many plagiarism detection systems have been developed and used to spot plagiarized texts. However, developing an efficient system requires a lot of plagiarized data to serve as examples for machine learning. This kind of data is also required as a standard test set for evaluating the performances of different plagiarism detection systems. In many languages, especially English, many plagiarism corpora have been created to meet these purposes (Clough & Stevenson, 2011; Mohtaj, Asghari, & Zarrabi, 2015; Potthast, Hagen, Völske, & Stein, 2013). However, to the best of our knowledge, no plagiarism corpus has been developed for a low-resource language like Thai. This study presents the first work on the development of a Thai plagiarism corpus.

In this paper, we describe the construction of an *Academic Thai Plagiarism Corpus (ATPC)*, a corpus that collects simulated academic plagiarism texts in Thai. Since authentic plagiarized texts are sparse and not well-organized to be used as training and/or testing data, many plagiarism corpora are designed and created to imitate plagiarism as much as possible. Usually, plagiarized texts are carefully created to reflect different linguistic mechanisms found in actual plagiarized texts, such as a corpus of plagiarized short answers (Clough & Stevenson, 2011). In order to be a benchmark in testing the performance of plagiarism detection systems, ATPC is designed to consist of two main types of texts: plagiarized texts and non-plagiarized texts. Plagiarized texts are categorized into four types based on the degree of linguistic mechanisms used in plagiarism.

As the first large plagiarism corpus in Thai, ATPC will be an important resource and benchmark for developing more effective plagiarism detection systems for Thai. Moreover, since some parts of ATPC are manually created by asking subjects to paraphrase/plagiarize a given text, the corpus will be useful not only for training and testing plagiarism detection systems, but also for studying linguistic strategies in paraphrasing/plagiarizing, which should be fundamental knowledge for plagiarism detection.

The rest of this paper presents theoretical frameworks and background that lead to the design of the plagiarism corpus. In a later section, the corpus creation processes are sequentially described. It then shows the properties of the created corpus, followed by data from the corpus analysis. At the end of this paper are conclusions and suggestions for future work.

LITERATURE REVIEW

Constructing a plagiarism corpus is related to various issues, including the definition of plagiarism, a taxonomy of plagiarism, and a framework for creating existing plagiarism corpora. In order to link the previous knowledge to this work and to set guidelines for creating a valid plagiarism corpus, relevant works **are compiled in the literature review section.**

DEFINITION OF PLAGIARISM

There is no general agreement on a standard definition of plagiarism to satisfy all situations (Sutherland-Smith, 2008, p. 57). The meaning of plagiarism depends on the interpretations in various contexts. However, based on previous literature, we can classify the definitions of plagiarism into three groups.

The first group defines plagiarism as the theft of another person's ideas, works, or writings. The definition in this respect considers academic work as a property that can be stolen and plagiarism as an act of stealing. This kind of definition appears in definitions given by Park (2003, p. 472), Ross and Thomas (2003, p. 211), and Sriganesh and Iyer (2007, p. 146). It also includes definitions that appear in a number of dictionaries, for example, *The Compact Edition of the Oxford English Dictionary* (Henry, 1971, p. 2192).

The second group defines plagiarism in terms of obscurity, cheating, or deception. In this group, plagiarism is considered an act of incorporating the ideas, writings, or words of others into the plagiarist's work by pretending or convincing readers that the work is their own. The definition of plagiarism in this respect is widely used in the field of education. It is similar to the first definition, but the intention of cheating is explicitly included in this definition; for example, the definitions given by Warn (2007, p. 196), Pecorari (2008, p. 4), Bretag and Mahmud (2009, p. 50), and Srisongkram (2011, p. 11).

The definition in the last group considers plagiarism as an inappropriate reuse of others' ideas, words, or writings**.** In this case, the phrase '*inappropriate reuse*' means using the information of others without showing the source or using works and thoughts of others without the usual acknowledgment**.** This definition is prevalent in later academic works, particularly in the field of information technology, since it does not require an interpretation of the intention behind the plagiarist's action. Such a definition appears in many works, for example, Sindhu.L, Thomas, and Idicula (2011, p. 65), Barrón-Cedeño, Vila, Martí, and Rosso (2013, p. 918), Ronald and Suharjito (2014, p. 168), and Mohtaj et al. (2015, p. 1).

In this study, we adopt the third definition because we think that it is difficult to interpret plagiarists' intentions**.** On the other hand, plagiarism should be defined based on textual characteristics, as it can be determined by linguistic and natural language processing methods**.** For these reasons, the definition of plagiarism we apply to this work is the adoption of others' ideas, words, or writings to present without showing the generally accepted acknowledgment.

CLASSIFICATION OF PLAGIARISM

To design a corpus that serves the purpose of training and testing plagiarism detection systems, we need to understand the types and patterns of plagiarism. Reviewing the works on plagiarism taxonomy allows us to create a variety of data that cover all types of actual plagiarism.

The classification of plagiarism began in the field of education, in order to study the methods of plagiarism and to enable the examiner to identify plagiarism in writings. The types of plagiarism classified by this viewpoint are often based on the writing strategies used to plagiarize and the intention behind the plagiarism; for instance, the classification proposed by Pecorari (2008, pp. 1-7), which divides textual plagiarism into two categories: *prototypical plagiarism* and *patchwriting*. Both these subcategories are distinguished by the presence or absence of the intention to deceive.

However, the classifications of plagiarism by the aforementioned viewpoint are a hassle for the examiner since these types often overlap (Bretag & Mahmud, 2009, p. 51), and proving the intention behind plagiarism is not easy. Therefore, when technology is advanced enough, the classifications should support a machine detection process.

When Clough and Stevenson (2009, 2011) created a corpus of plagiarized short answers, they classified the plagiarism in the corpus into four categories based on the levels of plagiarism: '*near copy*' for copy-and-paste from the original text, '*light revision*' for alteration of the original text by substituting words with synonyms and performing some grammatical changes, '*heavy revision*' for reorganizing original text at the syntactic level and paraphrasing, and '*non-plagiarism*' for texts that participants create based on their knowledge.

It is noteworthy that the classification in Clough and Stevenson's (2009, 2011) work is not based on the intention to deceive, but rather focuses on the methods used in plagiarism. Thus, this classification is suitable for computer-based processing, which corresponds to the purposes of this work.

There is another interesting classification proposed by Alzahrani, Salim and Abraham (2012), which is a new taxonomy of plagiarism based on the plagiarist's behavioral viewpoint. They also claim that their classification supports deep understanding of different linguistic patterns in committing plagiarism. The categories that appear in their work are divided into two main categories: '*literal plagiarism*' and '*intelligent plagiarism*'. Literal plagiarism is the simple and time-saving methods for plagiarizing, which consist of '*exact copy*' for copy-and-paste from the original text, '*near copy*' for inserting, deleting, or replacing words, including merging or splitting sentences, and '*modified copy*' for phrase reordering and syntactic editing. For intelligent plagiarism, they defined it as an attempt to deceive the readers into believing that the text is the plagiarist's own work. Intelligent plagiarists try to hide, obfuscate, and change the original text using various intelligent methods, including '*text manipulation*' which consists of paraphrasing and summarizing original text, '*translation*' both by humans and machines, and '*idea adoption*' which is considered to be the most intense plagiarism method and most difficult to detect.

Although intention is used as a categorization criterion, the classification shown in Alzahrani, Salim and Abraham's (2012) work provides details that link the linguistic mechanisms used in plagiarism. This issue corresponds to the findings of Barrón-Cedeño et al. (2013, p. 943) and Taerungruang and Aroonmanakun (2015, p. 62), which point out that there are linguistic mechanisms behind the plagiarism processes. However, some linguistic mechanisms that appear as methods of intelligent plagiarizing are quite complicated, as some methods are overlapping, so it may not be easy to apply in the corpus.

By considering the classifications proposed in Clough and Stevenson's work (2009, 2011) and Alzahrani, Salim and Abraham's (2012) work, we conclude that the types of plagiarism can be divided into different levels, and the application of higher level language mechanisms will result in a higher level of plagiarism type, which is more complex and different from the original text. With these conclusions, in our corpus, the data is divided into categories according to the level of plagiarism, which is due to the application of different levels of linguistic mechanisms. However, to suit the actual use, we omitted some details from the reviewed works. The types of data contained in the corpus are further detailed in the next section.

RELATED CORPORA

To design ATPC as a standard corpus for studying and testing plagiarism detection systems for Thai, previous frameworks used to create plagiarism corpora in other languages are reviewed. This section describes some works and designs of previous plagiarism corpora.

In the field of computer science and information, before the study of plagiarism was widely known, there was a growing emphasis on the study of text processing. In a way that resembles plagiarism, the term '*text reuse*' had been proposed. Due to interest in this field, as a benchmark corpus, *the METER corpus* (Gaizauskas et al., 2001) was created for the study of text reuse in journalism. This corpus consists of a set of news stories written by the Press Association (PA) and a set of stories about the same news events as published in nine British newspapers. Each story from the newspaper was assigned a level of text reuse at one of three levels: *Wholly Derived*, *Partially Derived*, and *Non-derived*, based on the amount of text of the same event written by the Press Association. Since METER was manually constructed by selecting texts and annotating types of reuse at the document-level, the corpus is a bit small. It contains 1,716 texts (over 500,000 words). Because of the way to determine the level of text reuse, in the later period, this corpus was also widely used for the evaluation of plagiarism detection systems (Cheema et al., 2015, p. 2).

In the later period, plagiarism was studied in various disciplines. One of the most well-developed fields is the detection of plagiarism; in this regard, many corpora have been created as benchmarks for measuring the effectiveness of detection methods. *The PAN-PC corpora* (Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011; Potthast, Stein, Barrón-Cedeño, & Rosso, 2010; Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009) are the series of English plagiarism corpora which are widely used in this field.

For constructing PAN-PC-10, Potthast et al. (2010, pp. 4-6) proposed three levels of plagiarism authenticity: '*real plagiarism*' for actual cases of plagiarism, '*simulated* *plagiarism*' for plagiarism cases manually imitated by humans, and '*artificial plagiarism*' for plagiarism cases algorithmically generated by machines. Based on this concept, they pointed out the limitations on collecting real plagiarism cases, which are the main reason that drives them to choose simulated and artificial plagiarism cases as data in their corpus. For this reason, PAN-PC-10 features various kinds of plagiarism cases, including obfuscated cases that have been generated automatically and manually. Since a large part of PAN-PC-10 was generated automatically, the corpus is larger than other plagiarism corpora. It contains 27,073 texts, in which 68,558 plagiarism cases are included.

Other corpora in PAN's series that were later created improved the proportion of data to fit into the competition between different detection systems. Nevertheless, the basic idea used to create such corpora is based on PAN-PC-10. For example, Mohtaj, Asghari and Zarrabi, (2015) proposed a new way to automatically create simulated plagiarism by measuring similarity between sentences.

In contrast to the PAN-PC corpora, another corpus composed of all simulated plagiarism cases was created (Clough & Stevenson, 2009, 2011). It consists of both plagiarized and non-plagiarized texts of between 200-300 words in English. However, the size of the corpus is rather small because all the texts were manually written by the subjects.

Plagiarism corpora in other languages were also created for the same purpose. For example, *UPPC* (Sharjeel, Rayson, & Nawab, 2016) is a plagiarism corpus of Urdu, which is manually generated for evaluating Urdu plagiarism detection systems. But no plagiarism corpus of Thai has been developed and released as a test set for evaluation.

Based on the review of related corpora, it can be concluded that the texts contained in the corpus should be subdivided into different levels of plagiarism. In keeping with the classification of plagiarism discussed in the previous section, we decided to classify the data in our corpus based on the linguistic mechanism used in plagiarism. The data can be generated by machines and simulated by humans. For details on creating our corpus, please refer to the next section.

CORPUS CONSTRUCTION PROCESS

In this section, we present the design and creation of our plagiarism corpus, ATPC, as a result of the compiled knowledge derived from reviewing the relevant works discussed in the previous section.

CORPUS DESIGN

In order to be a benchmark in testing the performance of the plagiarism detection system, our corpus, ATPC, is designed to include both plagiarized and non-plagiarized texts. It is composed of 91,250 texts, in which 12,500 plagiarism cases are included. Any Thai plagiarism detection system can search and identify plagiarized paragraphs in the corpus in a similar way to those tested on PAN-PC-10. Besides providing full texts for testing, ATPC also aligns source and suspicious texts. This makes it ready to be used for evaluating systems on measuring the degree of plagiarism, without concern over the retrieval part.

Based on the level of application of the linguistic mechanism in plagiarizing, plagiarized data is subdivided into four subcategories: '*Exact Copy (EC)*' for straightforward copying of original text without modification, '*Near Copy (NC)*' for editing original text at the word level, i.e. inserting or deleting words in the original text, '*Modified Copy (MO)*' for modifying original text at the syntactic level, i.e. inserting, deleting, or moving a phrase or clause in the original text, and '*Paraphrase (PA)*' for editing the original text at the semantic level, e.g. word substitution in terms of semantic relation, word conversion to synonymous phrase, or rewriting the sentence.

In terms of size and proportion of data, ATPC has 50,000 pairs of paragraphs, divided into 37,500 non-plagiaristic pairs of paragraphs (75% of data) and 12,500 plagiaristic pairs of paragraphs (25% of data). The 12,500 plagiaristic pairs of paragraphs are also divided according to the level of plagiarism as follows: exact copy 3,750 pairs (30% of plagiaristic data), near copy 3,750 pairs (30% of plagiaristic data), modified copy 3,750 pairs (30% of plagiaristic data), and paraphrase 1,250 pairs (10% of plagiaristic data).

DATA COLLECTION

All of the raw data used to create our corpus come from the Master's theses and Doctoral dissertations published in Thai, totaling 2,624 copies, which are sponsored by the Graduate School, Chulalongkorn University. All theses are classified into 2 groups according to their related fields: *science (SC)* and *humanities and social science (HS)*.

Thai word segmentation is applied in all theses. Then, paragraphs with the specified length are selected and divided into 3 groups: 50-100 words for short paragraphs, 101-150 words for medium paragraphs, and 151-200 words for long paragraphs. By this method, all 489,713 raw paragraphs are derived, as shown in Table 1.

TABLE 1. Number of raw paragraphs classified by related field and size

Field and size of raw paragraphs					
Science			Humanities and Social science		
Short	Medium	Long	Short	Medium	Long
93,500	49,108	23,856	180,896	95,452	46,901

SIMULATION OF PLAGIARISTIC DATA

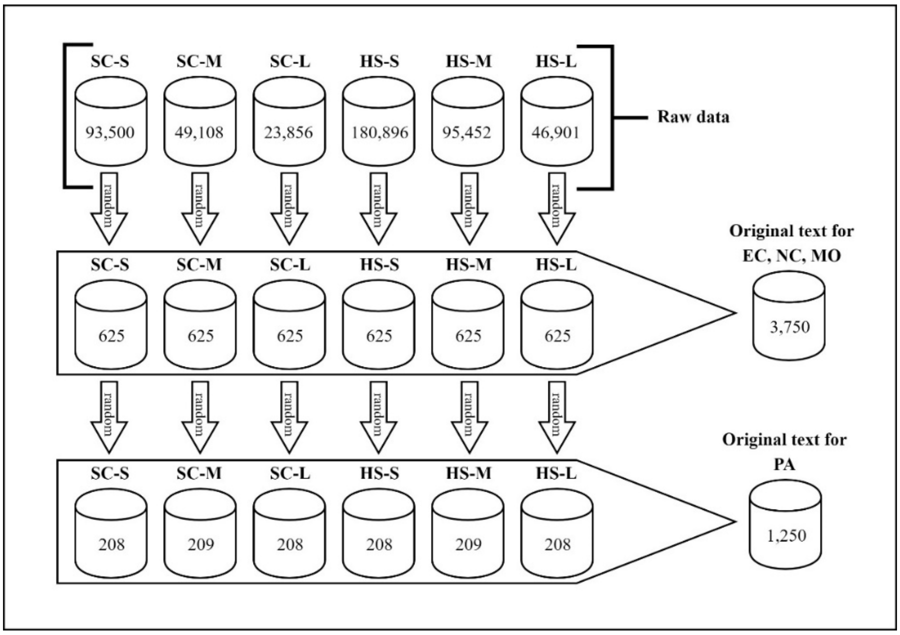


FIGURE 1. The process of obtaining the original text for simulating plagiarism

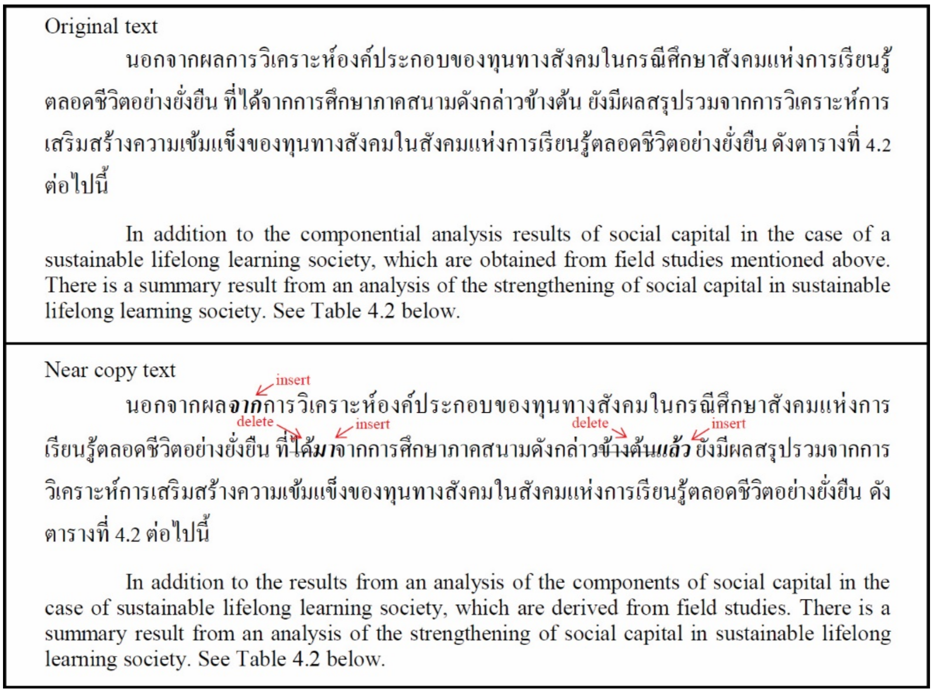
After obtaining the raw data as described in the previous subsection, we selected a number of paragraphs at random to serve as the original text for simulating plagiaristic data. As shown in Figure 1, for the proper distribution of data, each group of its field and size is randomly selected in 625 paragraphs to use as the original text for simulating EC, NC, and MO data. For PA data, each of the 625 paragraphs is randomly re-drawn to 208 or 209 for using as original text.

CREATING EXACT COPY DATA

The EC data are automatically generated by the machine. Since this kind of plagiarism does not need any linguistic knowledge, the algorithm used to simulate this part of the data is simple. By this concept, the creation of the EC involved just a copying and pasting the original text of all 3,750 paragraphs into plagiarized texts and arranging them in pairs of paragraphs as well.

CREATING NEAR COPY DATA

For the creation of data as NC plagiarized texts, which is the original text edited at the word level, we have provided word-modifying rules indicating a list of words that can be inserted or removed without altering the overall meaning of the text, along with the left and right context of those words. By automatically applying these rules, each word will be deleted or inserted or replaced with another word if its left and/or right contexts match those specified in the rules, when it is found in the original text. Figure 2 shows an example of a near copy text compared to the original text. From the example, it can be seen that all five words in the near copy text are inserted and deleted based on the word list prepared. This process is applied in all 3,750 original paragraphs, with the condition that in each paragraph at least one word is inserted or deleted.


Original text

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

In addition to the componential analysis results of social capital in the case of a sustainable lifelong learning society, which are obtained from field studies mentioned above. There is a summary result from an analysis of the strengthening of social capital in sustainable lifelong learning society. See Table 4.2 below.

Near copy text

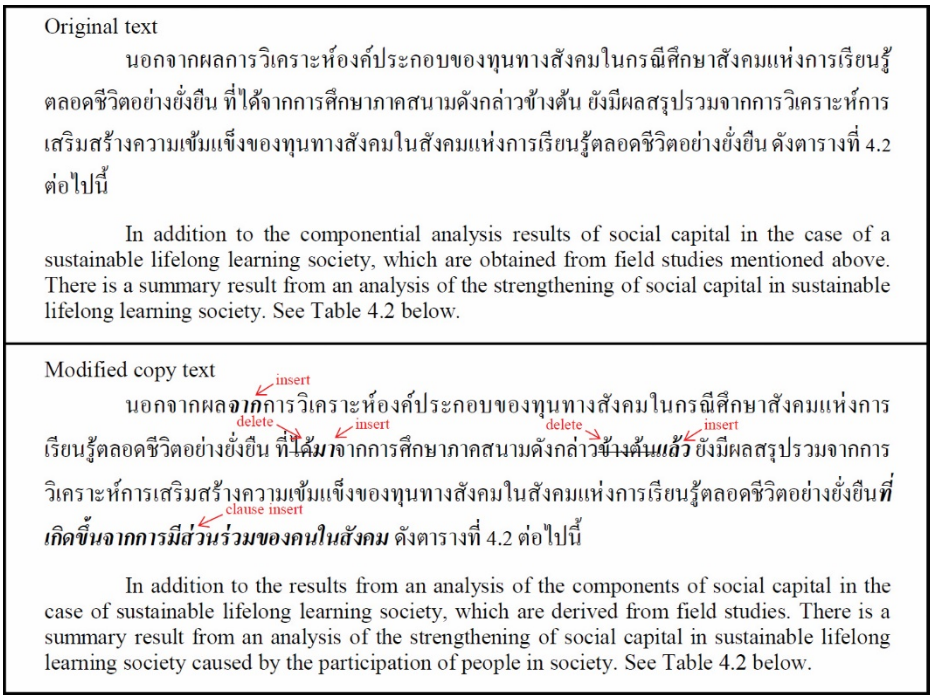
นอกจากผล***จาก***การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ~~ที่~~ได้***มา***จากการศึกษาภาคสนามดังกล่าว~~ข้างต้น~~***แล้ว*** ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

In addition to the results from an analysis of the components of social capital in the case of sustainable lifelong learning society, which are derived from field studies. There is a summary result from an analysis of the strengthening of social capital in sustainable lifelong learning society. See Table 4.2 below.

FIGURE 2. Example of near copy text compared to original text

CREATING MODIFIED COPY DATA

As for MO plagiarized tests, the data are manually created by the first author and three participants who have have completed their Master's degree in Thai and are currently teaching Thai academic writing related subjects at the university. The participants were asked to create data by inserting, deleting or moving a phrase or clause in the NC data without changing the main idea of the text. For a more in-depth understanding, the participants received a detailed guide on phrases and clauses in Thai, how to edit the text, and how to mark up editing information. Figure 3 shows an example of a modified copy text compared to the original text. From this example, it can be seen that at the end of the plagiarized text, what appears is an insertion of a subordinate clause.



Original text

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

In addition to the componential analysis results of social capital in the case of a sustainable lifelong learning society, which are obtained from field studies mentioned above. There is a summary result from an analysis of the strengthening of social capital in sustainable lifelong learning society. See Table 4.2 below.

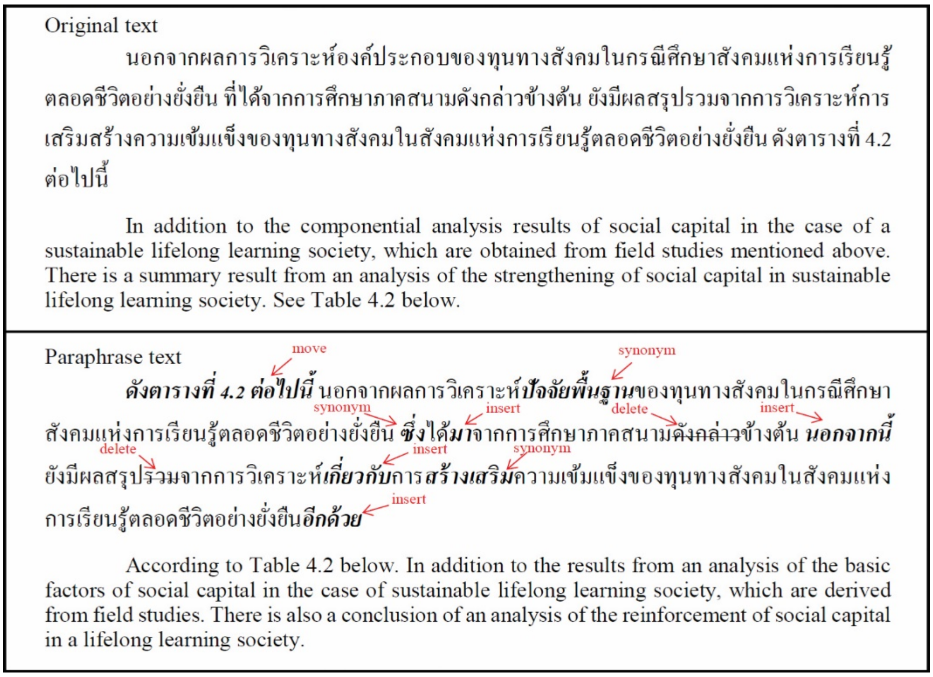
Modified copy text

นอกจากผล***จาก***การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่~~ได้~~***มา***จากการศึกษาภาคสนามดังกล่าว~~ข้างต้น~~***แล้ว*** ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน***ที่เกิดขึ้นจากการมีส่วนร่วมของคนในสังคม*** ดังตารางที่ 4.2 ต่อไปนี้

In addition to the results from an analysis of the components of social capital in the case of sustainable lifelong learning society, which are derived from field studies. There is a summary result from an analysis of the strengthening of social capital in sustainable lifelong learning society caused by the participation of people in society. See Table 4.2 below.

FIGURE 3. Example of modified copy text compared to original text

CREATING PARAPHRASE DATA



Original text

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

In addition to the componential analysis results of social capital in the case of a sustainable lifelong learning society, which are obtained from field studies mentioned above. There is a summary result from an analysis of the strengthening of social capital in sustainable lifelong learning society. See Table 4.2 below.

Paraphrase text

ดังตารางที่ 4.2 ต่อไปนี้ นอกจากผลการวิเคราะห์***ปัจจัยพื้นฐาน***ของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ***ซึ่ง***ได้***มา***จากการศึกษาภาคสนาม~~ดังกล่าว~~ข้างต้น ***นอกจากนี้*** ยังมีผลสรุป~~รวม~~จากการวิเคราะห์***เกี่ยวกับ***การ***สร้างเสริม***ความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน***อีกด้วย***

According to Table 4.2 below. In addition to the results from an analysis of the basic factors of social capital in the case of sustainable lifelong learning society, which are derived from field studies. There is also a conclusion of an analysis of the reinforcement of social capital in a lifelong learning society.

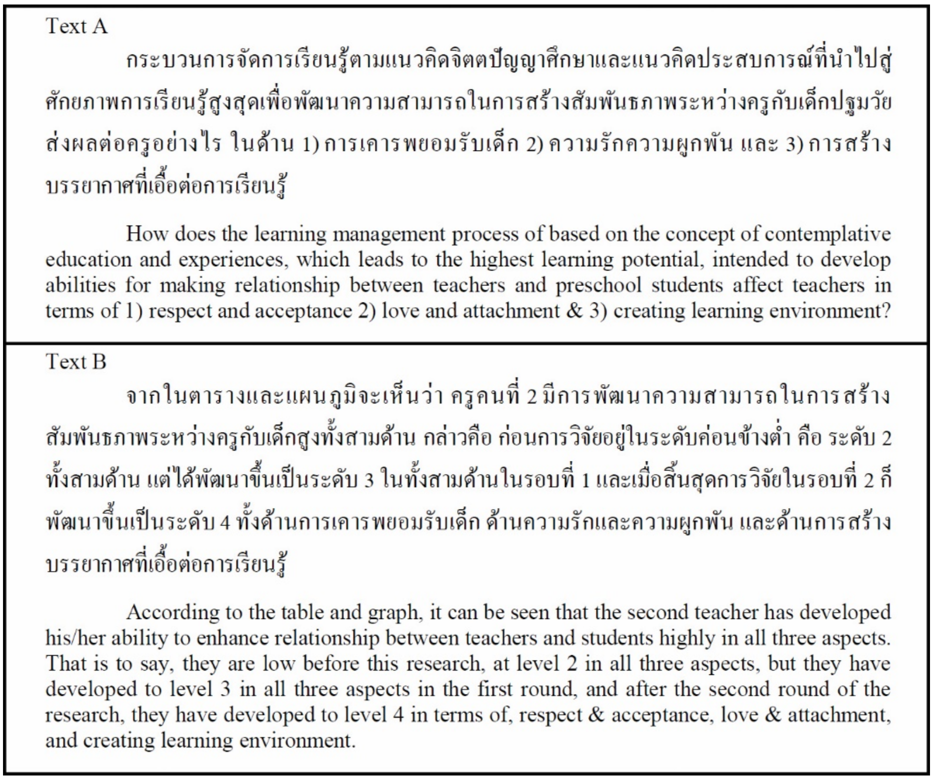
FIGURE 4. Example of paraphrase text compared to original text

Similar to the MO plagiarized texts, the PA plagiarized texts are manually created by three participants with the same qualifications as the participants who created the MO data. With a guide to edit the text, the participants were requested to paraphrase 1,250 original texts without changing the subject matter of the text. Figure 4 shows an example of a paraphrase text compared to the original text. In this example, it can be seen that the participant used a variety of linguistic mechanisms for paraphrasing, including insertion, deletion, synonym substitution, and moving a clause.

SIMULATION OF NON-PLAGIARISTIC DATA

Non-plagiarized texts are composed of two parts. The first part are those original texts used in the process of creating plagiarized texts mentioned above. The second part of original texts were selected based on the degree of similarity to plagiarized texts. Non-plagiarized texts are also used for training and testing the system's performance in identifying whether the input is plagiarized. Hence, to represent the suspicious data retrieved by the system, the pairs of paragraphs created for this purpose need to be similar on a certain level which can be suspected as plagiarism.

In keeping with the concepts discussed earlier, we designed a pair of paragraphs in this type of data to be derived from matching paragraphs with the same word at the specified level. For this purpose, the algorithm is designed to measure the similarity between all paragraphs in the same field and size. If there are any pairs of paragraphs with a tri-gram of word similarity value between 0.2 and 0.5 measured by using Sørensen-Dice coefficient, then the paragraphs will be selected.



Text A

กระบวนการจัดการเรียนรู้ตามแนวคิดจิตตปัญญาศึกษาและแนวคิดประสบการณ์ที่นำไปสู่ศักยภาพการเรียนรู้สูงสุดเพื่อพัฒนาความสามารถในการสร้างสัมพันธภาพระหว่างครูกับเด็กปฐมวัยส่งผลต่อครูอย่างไร ในด้าน 1) การเคารพยอมรับเด็ก 2) ความรักความผูกพัน และ 3) การสร้างบรรยากาศที่เอื้อต่อการเรียนรู้

How does the learning management process of based on the concept of contemplative education and experiences, which leads to the highest learning potential, intended to develop abilities for making relationship between teachers and preschool students affect teachers in terms of 1) respect and acceptance 2) love and attachment & 3) creating learning environment?

Text B

จากในตารางและแผนภูมิจะเห็นว่า ครูคนที่ 2 มีการพัฒนาความสามารถในการสร้างสัมพันธภาพระหว่างครูกับเด็กสูงทั้งสามด้าน กล่าวคือ ก่อนการวิจัยอยู่ในระดับค่อนข้างต่ำ คือ ระดับ 2 ทั้งสามด้าน แต่ได้พัฒนาขึ้นเป็นระดับ 3 ในทั้งสามด้านในรอบที่ 1 และเมื่อสิ้นสุดการวิจัยในรอบที่ 2 ก็พัฒนาขึ้นเป็นระดับ 4 ทั้งด้านการเคารพยอมรับเด็ก ด้านความรักและความผูกพัน และด้านการสร้างบรรยากาศที่เอื้อต่อการเรียนรู้

According to the table and graph, it can be seen that the second teacher has developed his/her ability to enhance relationship between teachers and students highly in all three aspects. That is to say, they are low before this research, at level 2 in all three aspects, but they have developed to level 3 in all three aspects in the first round, and after the second round of the research, they have developed to level 4 in terms of, respect & acceptance, love & attachment, and creating learning environment.

FIGURE 5. Example of non-plagiaristic pair of paragraphs

Figure 5 shows an example of a non-plagiaristic pair of paragraphs with word tri-gram similarity values of 0.23.Considering this example, it is found that the content of both paragraphs is very close. They both refer to the new educational process for early childhood.

All 37,500 paragraph pairs are selected by this algorithm, with the volume control of each field and size in a similar way.

CORPUS PROPERTIES

After creating the corpus according to the methods shown in the previous section, the properties that illustrate the overall nature of the corpus created is presented in this section.

The corpus contains 91,250 texts, in which 12,500 plagiarized cases are included. The corpus is also arranged into 50,000 pairs of paragraphs, which are subdivided into plagiarized and non-plagiarized data. According to the proportion of data that is defined and described in the previous section, Table 2 shows the amount of data contained in the corpus by type, field, and paragraph size.

TABLE 2. The number of pairs of paragraphs contained in the corpus

Data type	**Pair type**	**Field and size of paragraph pair**						**Total**
		Science			**Humanities and Social science**			
		S	**M**	**L**	**S**	**M**	**L**	
Plagiarized	SR-EC	625	625	625	625	625	625	3,750
	SR-NC	625	625	625	625	625	625	3,750
	SR-MO	625	625	625	625	625	625	3,750
	SR-PA	208	209	208	208	209	208	1,250
Non-plagiarized	NA-NB	7,194	7,194	4,362	6,250	6,250	6,250	37,500

For the number of words, the corpus contains 11,933,188 words (tokens) in total and 51,874 unique words. In Table 3, which shows the average number of words in each paragraph, it is noted that the plagiaristic paragraphs are shorter. This is in line with findings by Barrón-Cedeño et al. (2013, p. 943) and Taerungruang and Aroonmanakun (2015, p. 58), stating that the plagiarized text tends to be shorter than the original text. However, it can be seen that the NC data is longer than the original text. This may be due to the process of creating this type of text that inserts and deletes words in the original text based on the list of words prepared, i.e. the word list may cause the number of insertions to be greater than the number of deletions.

TABLE 3. Average number of words per paragraph

Data type	**Data sub-type**	**Average number of words**		
		S	**M**	**L**
original	$SR_{EC,NC,MO}$	75.01	122.10	173.18
	SR_{PA}	74.92	122.67	173.21
plagiarized	EC	75.01	122.10	173.18
	NC	75.78	123.27	174.41
	MO	72.60	118.16	167.68
	PA	73.64	116.98	164.52
non-plagiarized	NA	73.62	118.84	171.45
	NB	74.17	121.95	171.63

The corpus is saved in plain text (.txt) format, where each paragraph is saved as a file. The name of each file is set to indicate the metadata of a paragraph, and refers to another file that records its matched paragraph. And for the corpus annotation, in the beginning, word segmentation is applied in all paragraphs.

CORPUS ANALYSIS & VALIDATION

To illustrate the validity of the corpus created, in this section, we present the results of a corpus analysis. For this purpose, the concept behind the criteria used in the analysis is first described, followed by detailed presentation of analysis of the result.

CRITERIA FOR ANALYSIS: SIMILARITY MEASUREMENT

In terms of plagiarism detection, one of the most widely used approaches is to measure the similarity between the suspicious texts. In this work, we apply the concept of Clough and Stevenson (2011, p. 17), which uses similarity to determine and distinguish the various types of plagiarism data that are compiled in the corpus. If the results indicate that each type of data is strictly separate, it will be asserted that the proposed methods of creating data are efficient, i.e. they can simulate each type of data without overlapping. Based on this concept, here we apply two similarity computing methods: *Longest common subsequence* and *Sørensen-Dice coefficient*.

LONGEST COMMON SUBSEQUENCE

The longest common subsequence (LCS) is a simple concept that has been used extensively in the field of natural language processing. The basic idea of LCS is to compare the sequence of two strings from left to right. Based on this concept, the result, LCS, is the sequence of common elements such that no longer string is available.

For the application for similarity computing, Clough and Stevenson (2011, p. 17) proposed to normalize LCS by computing the LCS between two texts and then dividing by the length of one of the two texts. This gives a normalized LCS (LCS_{norm}) value, which is a scale in range 0 to 1. If the result is 0, the compared texts are not similar; conversely, if the result is 1, the compared texts are identical.

Different LCS_{norm} values will show different levels of editing from the original text. Based on this idea, this work uses the LCS_{norm} value to represent the level of text editing at the character level.

SØRENSEN-DICE COEFFICIENT

Sørensen-Dice coefficient (Dice, 1945; Sørensen, 1948) is a statistic used for comparing the similarity of two samples. Its basic concept is similar to *Jaccard coefficient*, which is widely used in the field of information retrieval, i.e. it derives from the number of shared members against the total number of members. But Sørensen-Dice coefficient reduces the effect of members sharing between samples; thus, it can be seen as a measure of similarity over sets.

This work uses Sørensen-Dice coefficient to measure similarity between texts at the word level. Given an *n*-gram of length *n*, $S(A, n)$ is the set of word *n*-grams for text A; $S(B, n)$ is the set of word *n*-grams for text B; the quotient of similarity between text A and text B, QS_n, is defined following Eq. 1. Like LCS_{norm}, QS_n ranges between 0 and 1, with 0 indicating that there is no set of word *n*-grams shared between text A and B, and 1 indicating that both texts are identical.

$$QSn2SA,nSB,nSA,nSB,n \qquad (1)$$

To cover the editing of original text in various sizes and to identify the similarity level of various plagiarism types contained in this corpus, in the analysis, we compare *n*-gram sets of lengths 1-5.

RESULTS OF THE CORPUS ANALYSIS

In this subsection, we present the detailed results from a corpus analysis. First, we present the results of similarity analysis at the character level using the normalized LCS value as a basis for computing. We then present the results of the similarity analysis at the word level, using the Sørensen-Dice coefficient. And the results of the comparison between the groups of data in the corpus are presented last.

SIMILARITY BETWEEN TEXTS AT CHARACTER LEVEL

To analyze the similarity between texts at the character level, we compute the similarity between all the matched paragraphs in the corpus using $\mathrm{LCS}_{\mathrm{norm}}$ as an indicator. Table 4 shows the statistics as a result of computing the similarity between texts.

Considering the mean of $\mathrm{LCS}_{\mathrm{norm}}$, it can be seen that in the group of plagiarized data, the average has decreased gradually from SR-EC to SR-PA. This is consistent with the conclusion in the literature review that low-level plagiarism is more easily detected. As can be seen, plagiarized data type EC, a copy without changing the original text, has an average of 1, while the PA type, a paraphrase of the original text, has an average of 0.89. In contrast to plagiarized data, non-plagiarized data returns an average of only 0.53, indicating that only a fraction of the paragraphs in this data type are similar.

However, it is noted that the mean in each group of plagiarized data is very similar. This led to the observation that the use of character-level detection techniques may not be sufficient for actual use, especially when non-plagiarized data is alike in similarity to plagiarized data. In light of this observation, in the case of using this corpus as a benchmark, systems that rely on simple character-based detection techniques may not suffice.

TABLE 4. Statistics of similarity between texts at character level

Data type	Pair type	$\mathrm{LCS}_{\mathrm{norm}}$	
		Mean	SD
Plagiarized	SR-EC	1.0000	0.0000
	SR-NC	0.9788	0.1616
	SR-MO	0.9493	0.4909
	SR-PA	0.8932	0.0741
Non-plagiarized	NA-NB	0.5298	0.1068

SIMILARITY BETWEEN TEXTS AT WORD LEVEL

In this analysis, we measure the similarity between all the matched paragraphs at the word level, using the Sørensen-Dice coefficient as an indicator. Table 5 shows the statistics of similarity between various types of data, which are computed at the length of word *n*-gram from 1 (QS_1) to 5 words (QS_5).

Like similarity at the character level, the degree of similarity between the paragraphs is lower as the level of plagiarism increases. This indicates that high-level techniques of plagiarism tend to disintegrate the sequences of words in longer lengths than low-level techniques. On the other hand, the similarity values of plagiarized data are distinctly different from non-plagiarized data's.

It is also found that when the length of word *n*-gram is increased, the similarity decreases. Consider an average of QS_5, which is computed from the similarity of word five-grams; it is found that the similarity between each type of plagiaristic data is considerably different compared to QS_1. This leads to the observation that more sophisticated detection systems may return better results in the case of using this corpus as a benchmark.

TABLE 5. Statistics of similarity between texts at word level

Data type	Pair type	Statistics	QS_1	QS_2	QS_3	QS_4	QS_5
Plagiarized	SR-EC	Mean	1.0000	1.0000	1.0000	1.0000	1.0000
		SD	0.0000	0.0000	0.0000	0.0000	0.0000
	SR-NC	Mean	0.9660	0.9259	0.8871	0.8500	0.8145
		SD	0.2108	0.0416	0.0620	0.0890	0.0162
	SR-MO	Mean	0.9220	0.8731	0.8272	0.7838	0.7425
		SD	0.0564	0.0658	0.0787	0.0926	0.1065
	SR-PA	Mean	0.8808	0.7691	0.6839	0.6124	0.5510
		SD	0.0727	0.1225	0.1554	0.1784	0.1948
Non-plagiarized	NA-NB	Mean	0.5598	0.3830	0.3025	0.2458	0.2029
		SD	0.0884	0.0870	0.0809	0.0811	0.0834

COMPARING VARIOUS TYPES OF DATA

As mentioned at the beginning of this section, although the similarity of each data type is different, this cannot confirm that each type of data is not overlapping. The overlapping of data may result in the failure of the detection system testing, i.e., the system may not be able to make precise data classification decisions. For this reason, this final analysis is to test whether each type of data in our corpus is statistically significantly different.

In general, testing whether two or more samples are different can be done by comparing the average of each sample, known as one-way analysis of variance (one-way ANOVA), in which case the average of the similarity may be tested. However, using the average we have in this test is a violation of the prerequisites of ANOVA, because there is a sample in which the populations are not equal in variance. This is because the similarity value of all the members of SR-EC is always 1. Therefore, the test with this statistic cannot be applied.

For the reason mentioned above, we choose the *Kruskal-Wallis test* (Kruskal & Wallis, 1952) instead. The Kruskal-Wallis test, also called one-way ANOVA on ranks, is a rank-based non-parametric test used for comparing two or more independent samples of equal or different sample sizes. It is considered the non-parametric alternative to one-way ANOVA.

By the means mentioned above, each group of similarities measured from pairs of paragraphs, i.e. LCS_{norm}, QS_1, QS_2, QS_3, QS_4, and QS_5, was tested individually. The results of this test showed that differences between each data type are all significant (with Bonferroni post-hoc test, $p < 0.05$). Figure 6 shows the distribution of data in the corpus at various levels of similarity.

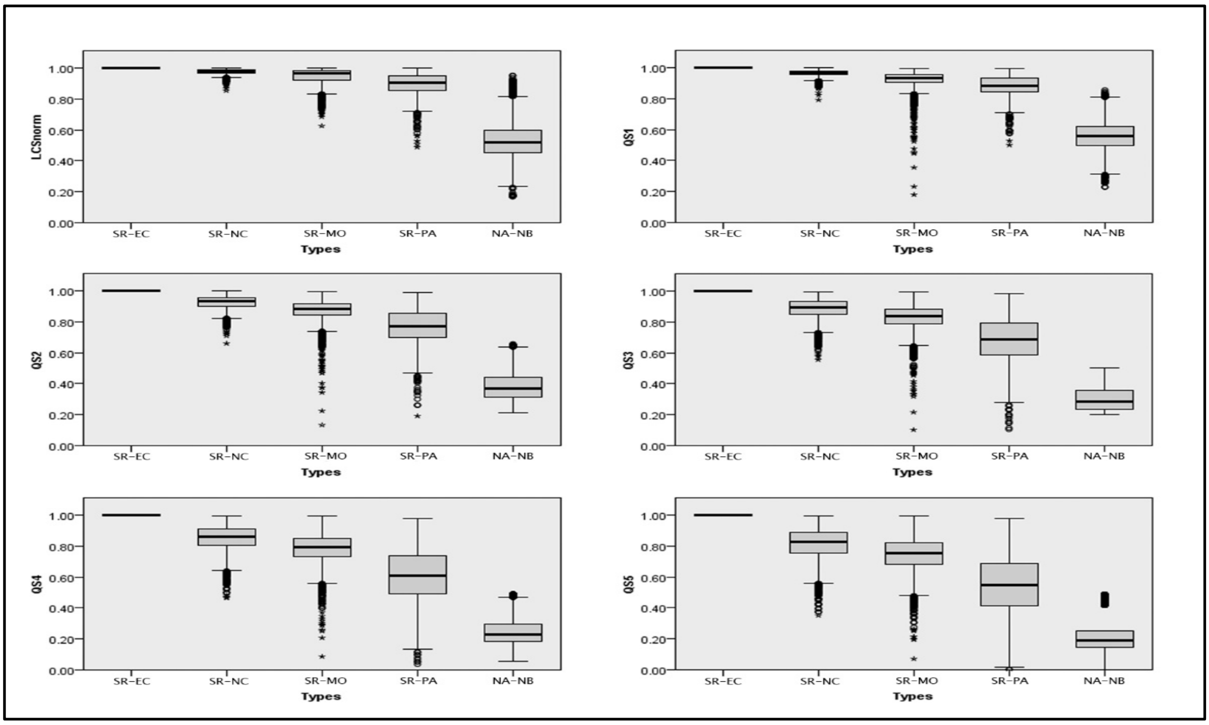


FIGURE 6. The distribution of data in the corpus at various levels of similarity

The results confirm that the methods used to create each type of data and to construct this corpus are effective. They can simulate plagiarized data and non-plagiarized data that is unique and does not overlap. Consequently, our corpus built with these methods can be used to train and test plagiarism detection systems efficiently.

CONCLUSION AND FUTURE WORK

The background and necessity of creating an academic Thai plagiarism corpus is described in the present study, with the purpose of presenting the process of designing the corpus and simulating each type of data for compiling the corpus.

For creation of data, we divided the data into various categories according to the degree of the linguistic mechanisms used in plagiarism. Data compiled in the corpus comes in two main ways: human creation and algorithmic generation by machine.

After discussing the data simulation, three steps of analysis to validate the corpus that was created, was proposed: a measurement of similarity between suspicious texts at the character level, a measurement of similarity between suspicious texts at the word level, and a comparison of different types of data compiled in the corpus based on the similarity measured.

The results of the analyses indicate that the corpus created by the proposed methods is effective in training and testing plagiarism detection systems. Since ATPC is designed as aligned paragraph-pairs between source and suspicious texts, those who want to evaluate their plagiarism detection systems can use this corpus to test their comparison component, in which classification of their systems as plagiarized or non-plagiarized will be compared to the answer key.

However, the corpus should be developed more efficiently. As well as our archives, there are other aspects to develop further, as follows.

In terms of corpus size, it is generally known that the larger the corpus, the higher quality of the work involved. An academic plagiarism corpus, like ours, should be extended. Since information in the academic world is constantly updated, if it is a static corpus, then it may not be useful in the future. Therefore, the amount of data in the corpus should be increased regularly. This may be done by adding various types of academic texts, such as academic articles, research articles, texts from text books or encyclopedias. In addition, it can be done by collaborating with other institutions to share original text for simulating plagiarized text.

In terms of academic discipline, the data in the corpus should be classified in more detail. This will help us know the balance of the data, which will lead to corpus improvement to cover a wide variety of disciplines. And it may also lead to the development of a specialized plagiarism corpus.

In terms of corpus annotation, for better detection performance, this corpus should be annotated more with various information, especially linguistic information; For examples, part-of-speech, boundary of clause, dependency relation in sentence, semantic role, discourse relation.

Finally, in terms of data quality, in case of more data, the quality of each type of data should be tested in various different ways. For example, using a machine learning model, e.g. Naive Bayes classifier, *k*-means clustering, to classify types of data in the corpus. This method will test the data and also be a method that is similar to that used in the detection system.

ACKNOWLEDGEMENT

This work is a part of the first author's dissertation. It is partially supported by the 90th anniversary of Chulalongkorn University fund (Ratchadaphiseksomphot endowment fund) of Graduate school, Chulalongkorn University (GCUGR1125592012D).

REFERENCES

Alzahrani, S. M., Salim, N. & Abraham, A. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). Vol. 42*(2), 133-149. http:dx.doi.org/10.1109/TSMCC.2011.2134847

Barrón-Cedeño, A., Vila, M., Martí, M. A. & Rosso, P. (2013). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics. Vol. 39*(4), 917-947.

Bretag, T. & Mahmud, S. (2009). A Model for Determining Student Plagiarism: Electronic Detection and Academic Judgement. *Journal of University Teaching & Learning Practice. Vol. 6*(1), 49-60.

Cheema, W. A., Najib, F., Ahmed, S., Bukhari, S. H., Sittar, A. & Nawab, R. M. A. (2015). A Corpus for Analyzing Text Reuse by People of Different Groups—Notebook for PAN at CLEF 2015. Paper presented at the CLEF 2015 Evaluation Labs and Workshop. A Conference at the Météo-CERFACS center. France, september.

Chulalongkorn University. (2012). *Academic Plagiarism: an Issue We Should Be Aware of.* Bangkok: Author.

Clough, P. & Stevenson, M. (2009). *Creating a corpus of plagiarised academic texts*. Paper presented at the Corpus Linguistics Conference (CL2009). UK: University of Liverpool, July.

Clough, P. & Stevenson, M. (2011). Developing a Corpus of Plagiarised Short Answers. *Language Resources and Evaluation. Vol. 45*(1), 5-24.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology. Vol. 26*(3), 297-302. http:dx.doi.org/10.2307/1932409

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P. & Piao, S. (2001). The METER Corpus: A corpus for analysing journalistic text reuse. Paper presented at the Corpus Linguistics 2001 conference. UK: Lancaster University, March.

Henry, J. A. (Ed.) (1971). *The Compact Edition of the Oxford English Dictionary*. Oxford: Oxford University Press.

Kruskal, W. H. & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association. Vol. 47*(260), 583-621. http:dx.doi.org/10.1080/01621459.1952.10483441

Mohtaj, S., Asghari, H. & Zarrabi, V. (2015). Developing monolingual English corpus for plagiarism detection using human annotated paraphrase corpus. Paper presented at the Conference and Labs of the Evaluation Forum (CLEF 2015) A Conference at the Météo-CERFACS center. France, september.

Park, C. (2003). In Other (People's) Words: Plagiarism by university students--literature and lessons. *Assessment & Evaluation in Higher Education, Vol. 28*(5), 471-488. http:dx.doi.org/10.1080/02602930301677

Pecorari, D. (2008). *Academic Writing and Plagiarism: A Linguistic Analysis*. London: Continuum.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2011). Overview of the 3rd international competition on plagiarism detection. Paper presented at the CLEF 2011 Labs and Workshops. A Conference at the Casa 400 Hotel. Netherlands: University of Amsterdam, September.

Potthast, M., Hagen, M., Völske, M. & Stein, B. (2013). Crowdsourcing interaction logs to understand text reuse from the web. Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). Bulgaria, August.

Potthast, M., Stein, B., Barrón-Cedeño, A. & Rosso, P. (2010). An evaluation framework for plagiarism detection. Paper presented at the 23rd International Conference on Computational Linguistics (COLING 2010). China, August.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A. & Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pp. 1-9). Valencia, Spain: CEUR-WS.org.

Ronald, A. & Suharjito. (2014). P Lagiarism Detection Algorithm Using Natural Language Processing Based on Grammar Analyzing. *Journal of Theoretical and Applied Information Technology. Vol. 63*(1), 168-180.

Ross, C. & Thomas, A. (2003). *Writing for Real: A Handbook for Writers in Community Service*. New York: Longman.

Sharjeel, M., Rayson, P. & Nawab, R. M. A. (2016). UPPC - Urdu paraphrase plagiarism corpus. Paper presented at the Language Resource and Evaluation Conference (LREC) 2016. Slovenia, May.

Sindhu.L, Thomas, B. B. & Idicula, S. M. (2011). A Study of Plagiarism Detection Tools and Technologies. *IJART. Vol. 1*(1), 64-70.

Sørensen, T. J. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons. *Kongelige Danske Videnskabernes Selskab. Vol. 5*(4), 1-34.

Sriganesh, V. & Iyer, P. (2007). Plagiarism and Medical Writing. *Indian Journal of Radiology and Imaging. Vol. 17*(3), 146-147. http:dx.doi.org/10.4103/0971-3026.34716

Srisongkram, W. (2011). *Development of plagiarism understanding of undergraduate students based on survey research and documentary analysis results.* Ph.D thesis, Chulalongkorn University, Bangkok, Thailand.

Sutherland-Smith, W. (2008). *Plagiarism, the Internet, and Student Learning: Improving Academic Integrity*. New York: Routledge.

Taerungruang, S. & Aroonmanakun, W. (2015). Konlawithii Laklok Ngaan Wichakan Phasa Thai: Kan Wikrao Thaang Phasasaat. [Plagiarism Strategies in Thai Academic Texts: a Linguistic Analysis]. *Language and Linguistics. Vol. 34*(1), 38-65.

Warn, J. (2007). Plagiarism Software: No Magic Bullet! *Higher Education Research & Development. Vol. 25*(2), 195-208. http:dx.doi.org/10.1080/07294360600610438

ABOUT THE AUTHORS

Supawat Taerungruang is a Ph.D. candidate in Linguistics at the Faculty of Arts, Chulalongkorn University. He is currently working on a dissertation about building plagiarism detection system using machine learning techniques and measuring the similarity of text.

Wirote Aroonmanakun is an associate professor at the Department of Linguistics, Faculty of Arts, Chulalongkorn University. His main research interests include the creation of corpora for linguistic use and natural language processing, and the development of basic tools for Thai language processing.