

# Design and Development of a Plagiarism Corpus in Thai for Plagiarism Detection

Santipong Thaiprayoon  
Pornpimon Palingoon  
Kanokorn Trakultaweekoon

Speech and Text Understanding (STU)  
Artificial Intelligence Research Group (AINRG)  
National Electronics and Computer Technology Center (NECTEC)

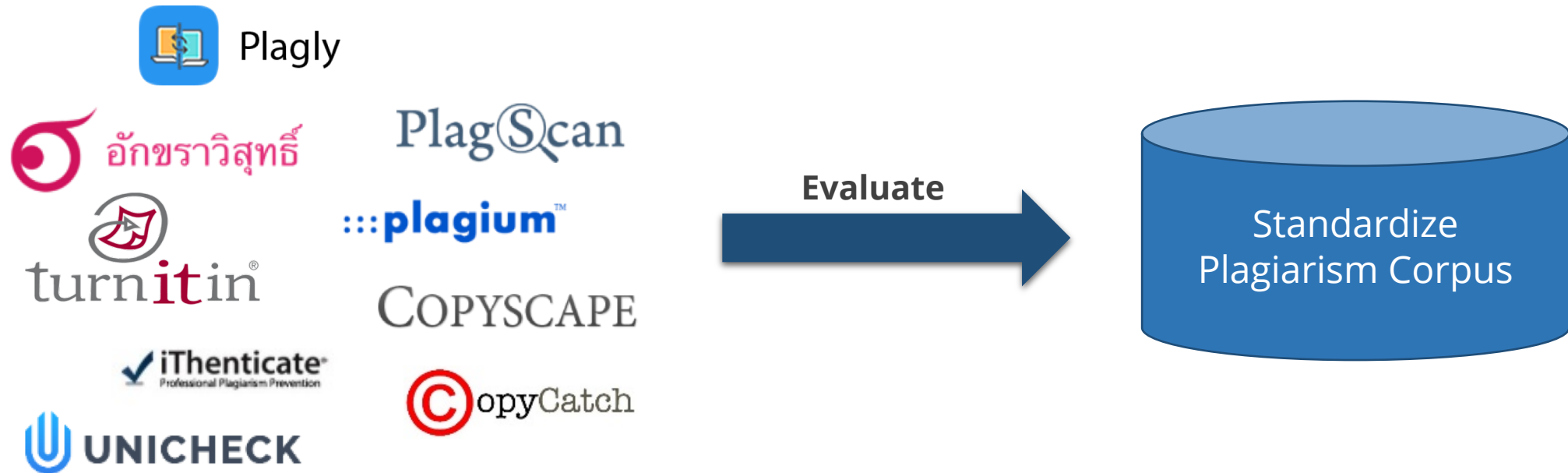


# Outline

- 1 Introduction and problems
- 2 Thai plagiarism construction
- 3 Plagiarism annotation tool
- 4 Corpus statistics and examples
- 5 Conclusions and future work

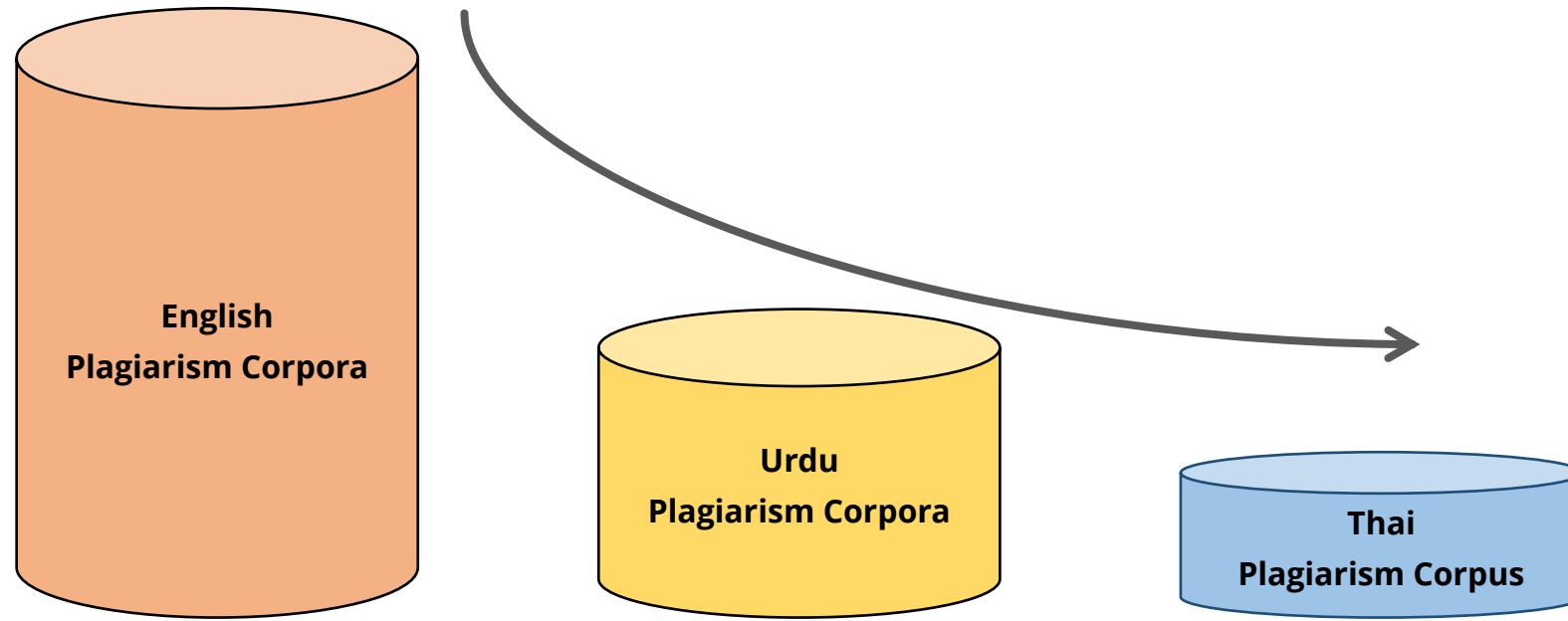
# Problems

- Plagiarism detection tools need a **benchmark corpus** containing real examples of plagiarized texts.



# Problems

- The plagiarism corpus in Thai is a **lack of a standardized resources**.



# Corpus construction approach



**Artificial plagiarism**

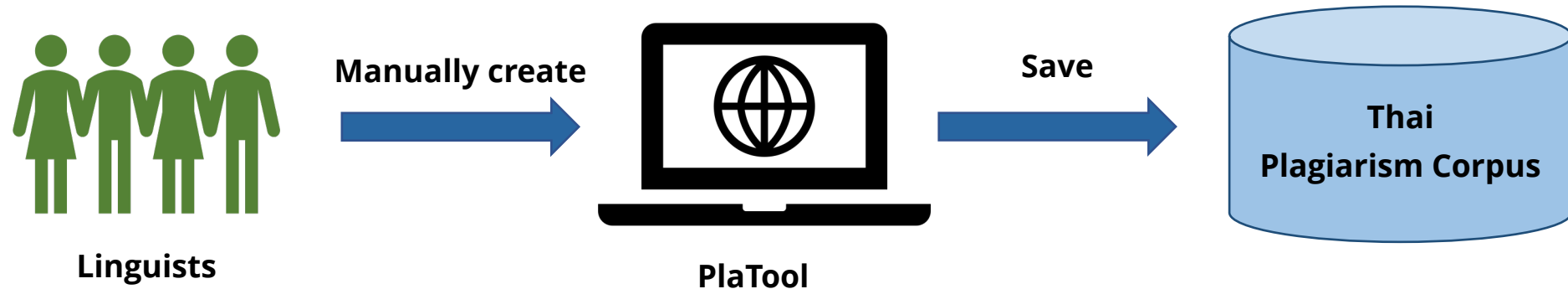
(Automatic)



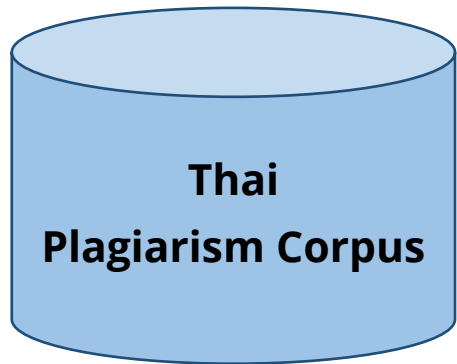
**Simulated plagiarism**

(Manual)

# Building the Thai plagiarism corpus

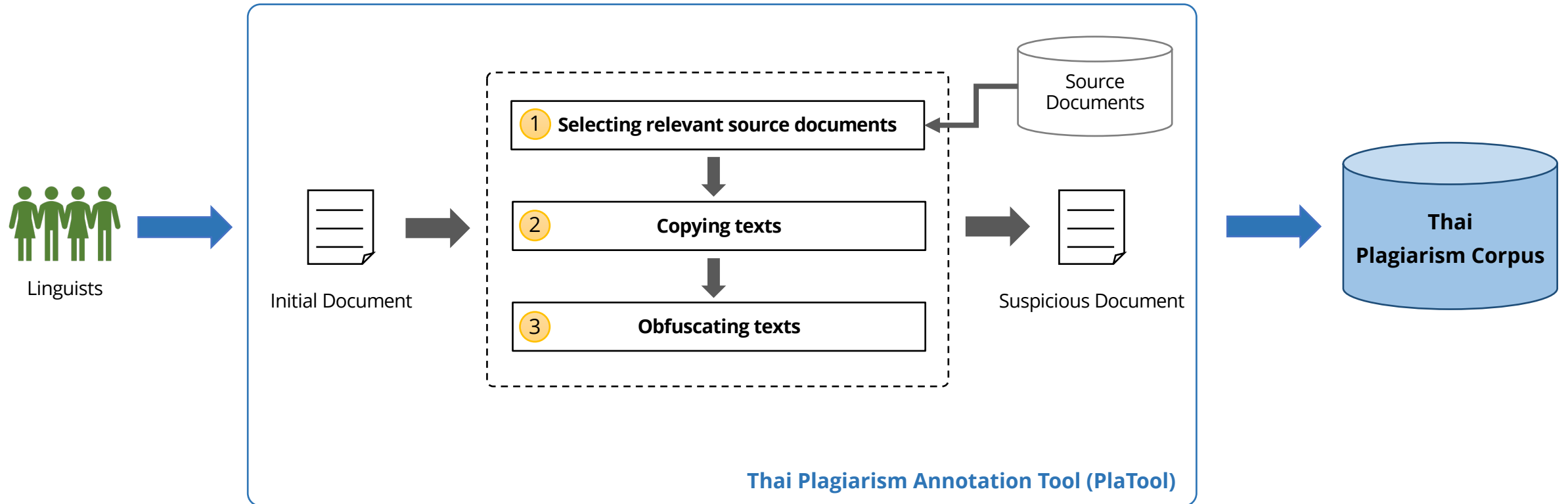


# Thai plagiarism classes




<b>Copy-based change</b>	This class is word-by-word copying that a text passage is copied from a set of source documents without any modifications and pasted into a document.
<b>Lexicon-based change</b>	A document is created from document sources by rewriting. This class contains three plagiarism cases, namely, lexical substitution, addition, and deletion.
<b>Structure-based change</b>	This class contains two main changes, namely, syntactic and discourse changes, which establish five plagiarism cases.
<b>Semantic-based change</b>	This class is about copying the idea or concept such as the result of experiment, summary or discussion from other studies in the same discipline.


# Overview of corpus construction





# Thai plagiarism annotation tool

 PlaTool (Thai Plagiarism Tagging tool)

 Suspicious

1 read and mark the beginning position

ขนมปัง

จากกรีกไปโรมถึงยุโรปตอนกลาง ศิลปะการทำขนมดำเนินไปอย่างเชื่องช้า แต่ได้ผลคงที่ ความเจริญก้าวหน้าอย่างมหาศาลทางด้านวิทยาศาสตร์และเทคโนโลยี ได้ทำให้เกิดวิวัฒนาการ อย่างใหญ่หลวงแก่การทำขนมอบในปัจจุบัน พื้นฐานของวิทยา

2 ประการ คือ ในกลางปี 1800 ได้มีการแนะนำเกี่ยวกับโรงไม้แปรรูป และในตอนที่ปลายศตวรรษนั้นได้มีการใช้ยีสต์ ซึ่งเป็น

ขึ้นฟู และมีการใช้กันอย่างแพร่หลาย และแป้งอีกชนิดหนึ่งที่ใช้ทำขนมปัง ได้แก่ แป้งข้าวโพด ซึ่งเป็นแป้งสกัดจากเมล็ดข้าวโพด เป็นผงสีขาวเหลืองนวล เนื้อเนียนละเอียดลื่นมือ เมื่อทำให้สุกมีลักษณะขุ่นและใส ไม่คืนตัวง่าย เมื่อตัวแป้งแห้งจะอยู่ตัวจับเป็นก้อนกรอบว่นเป็นมันวาว นิยมใช้ผสมเบเกอรี่เพื่อเพิ่มความนุ่มและความยืดหยุ่นให้กับเค้กบางชนิดเล็กน้อย เช่น แยมโรล เค้ก ฯลฯ ส่วนการผสมแป้งข้าวโพดในคุกกี้จะทำให้คุกกี้กรอบว่นยิ่งขึ้น

Fragment 1: mixed plagiarism cases

ส่วนผสมที่สำคัญ


ส่วนผสมที่สำคัญของขนมปังมีดังนี้

ข้าวสาลี : ขนมปังเกิดจากโปรตีนของแป้งสาลีที่มีชื่อว่า กลูเตน

Fragment 2: copy and paste

กลูเตน (มาจากภาษาละติน *gluten* แปลว่า กาว) กลูเตนเป็นโปรตีนที่สร้างความเหนียวให้กับก้อนแป้ง เกิดมาจากการรวมตัวของกลูเตนิน และ โกลอะดิน โดยพันธะไดซัลไฟด์ มีลักษณะเหนียว ยืดหยุ่น และไม่ละลายน้ำ โปรตีนชนิดนี้มีอยู่สูงในข้าวสาลี ในขณะที่ข้าวเจ้าที่คนไทยรับประทานในทุกวันนี้มีกลูเตนอยู่น้อยมาก นี่จึงเป็นสาเหตุที่ว่าทำไม ข้าวเจ้าจึงไม่สามารถนำมาทำขนมปังได้

2 search and select a related topic

Source แป้งทำขนม <https://cooking.kapook.com/view113107.html>  Add New

3 select the fragment

แป้งชนิดนี้ไม่ค่อยเหมาะในการทำเค้กที่ต้องการความนุ่มนวล เพราะหากนำมาทำเนื้อเค้กที่ได้จะแน่นกว่า แต่ก็มีบางสูตรใช้แป้งชนิดนี้แทนแป้งเค้ก ทำพวกเค้กที่ต้องการความนุ่มนวลก็ใช้ได้ แป้งชนิดนี้ใช้แทนแป้งเค้กได้มาจากกรรมวิธีการผลิตข้าวโพด ลักษณะแป้งที่ได้เป็นผงสีขาวเหลืองนวล เมื่อใช้มือสัมผัสเนื้อแป้งเนียนละเอียดลื่นมือ แต่เมื่อทำให้สุกมีลักษณะขุ่นและใส ไม่คืนตัวง่าย พอตัวแป้งแห้งจะอยู่ตัวจับเป็นก้อนกรอบว่นเป็นมันวาว สำหรับกรรมวิธีในเบเกอรี่นิยมใช้ผสมเบเกอรี่เพื่อเพิ่มความนุ่มและความยืดหยุ่นให้กับเค้กบางชนิดเพียงเล็กน้อย จะไม่ใช่เป็นส่วนผสมหลักสักเท่าไร เช่น แยมโรล เค้ก ฯลฯ ส่วนการผสมแป้งข้าวโพดในส่วนผสมคุกกี้จะทำให้คุกกี้กรอบว่นยิ่งขึ้น

4 modify the fragment

และแป้งอีกชนิดหนึ่งที่ใช้ทำขนมปัง ได้แก่ แป้งข้าวโพด แป้งชนิดนี้เป็นแป้งที่ได้มาจากกรรมวิธีการผลิตข้าวโพด ลักษณะแป้งที่ได้เป็นผงสีขาวเหลืองนวล เมื่อใช้มือสัมผัสเนื้อแป้งเนียนละเอียดลื่นมือ แต่เมื่อทำให้สุกมีลักษณะขุ่นและใส ไม่คืนตัวง่าย พอตัวแป้งแห้งจะอยู่ตัวจับเป็นก้อนกรอบว่นเป็นมันวาว สำหรับกรรมวิธีในเบเกอรี่นิยมใช้ผสมเบเกอรี่เพื่อเพิ่มความนุ่มและความยืดหยุ่นให้กับเค้กบางชนิดเพียงเล็กน้อย จะไม่ใช่เป็นส่วนผสมหลักสักเท่าไร เช่น แยมโรล เค้ก ฯลฯ ส่วนการผสมแป้งข้าวโพดในส่วนผสมคุกกี้จะทำให้คุกกี้กรอบว่นยิ่งขึ้น

5 insert modified fragment into suspicious text

✓ Insert Modified Text Reset

Symbol => Text : Inserting , Text : Removing , Text : Replacing

# A suspicious document format

```
<document language="th" reference="suspicious-document-129.txt">  
  <plagiarized source_reference="23428.txt" source_offset="121" source_length="251" this_offset="1265" this_length="220" />  
  <plagiarized source_reference="10257.txt" source_offset="502" source_length="247" this_offset="2351" this_length="217" />  
</document>
```

Source document	
<b>source_reference</b>	Name of a source document
<b>source_offset</b>	The position of first character corresponding to a source fragment in the source document
<b>source_length</b>	The number of characters in a source fragment of the source document

Suspicious document	
<b>this_offset</b>	The position of the first character corresponding to a plagiarized text in the suspicious document
<b>this_length</b>	The number of characters in a source fragment of the source document

# Corpus statistics

- We collected 111,000 articles from Thai Wikipedia and web pages.



109,200 articles



800 pages

# Corpus statistics

- The corpus consists of a set of suspicious documents and a set of source documents.
- A suspicious document may contain plagiarized passages from one or more source documents.

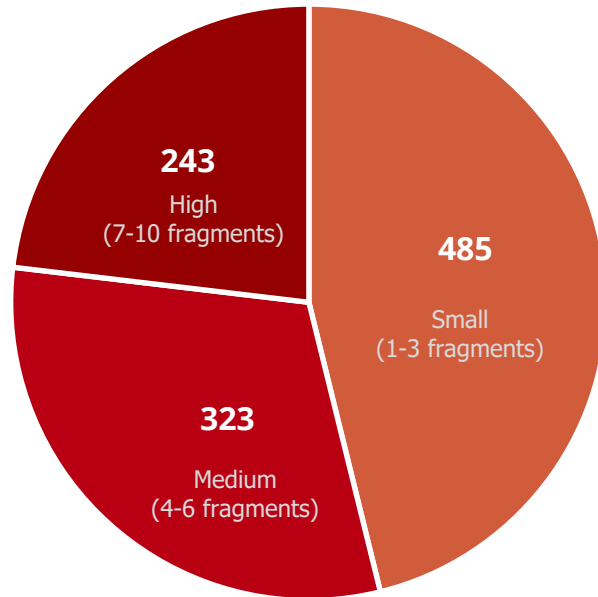


111,000 source documents

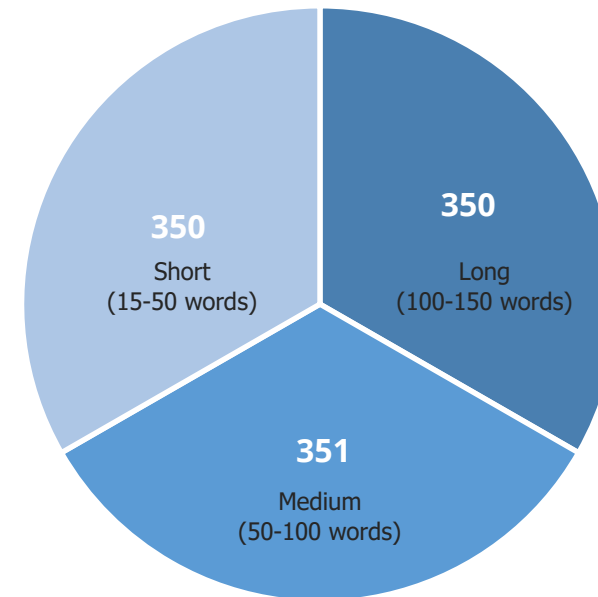


1,051 suspicious documents

# Plagiarism cases statistics

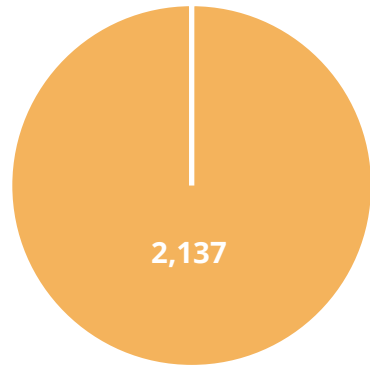


Number of Plagiarized Fragments

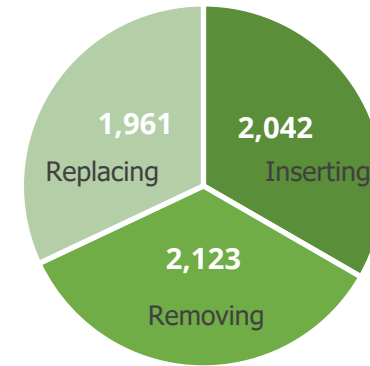


Plagiarized Fragment Lengths

# Statistics of Thai plagiarism classes

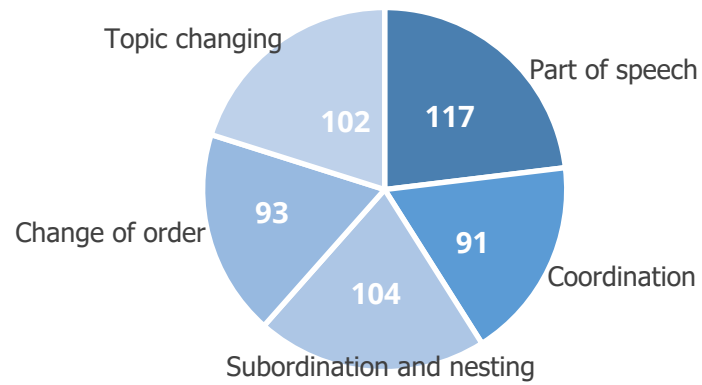


Copy-Based Change

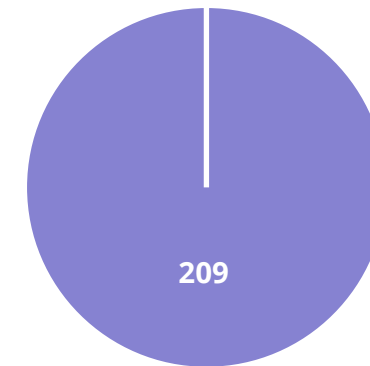


Lexicon-Based Change

Plagiarized cases : 8,979 cases



Structure-Based Change



Semantic-Based Change

# An example

## Example of lexicon-based change (Addition)

### In Thai

#### Original document

เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรมประเภทหนึ่ง

#### Modified document

เครื่องคอมพิวเตอร์ประกอบด้วยส่วนประมวลผล โดยทั่วไปเป็นหน่วยประมวลผลกลาง และแรม (หน่วยความจำ) ประเภทหนึ่ง

### Translation

#### Original document

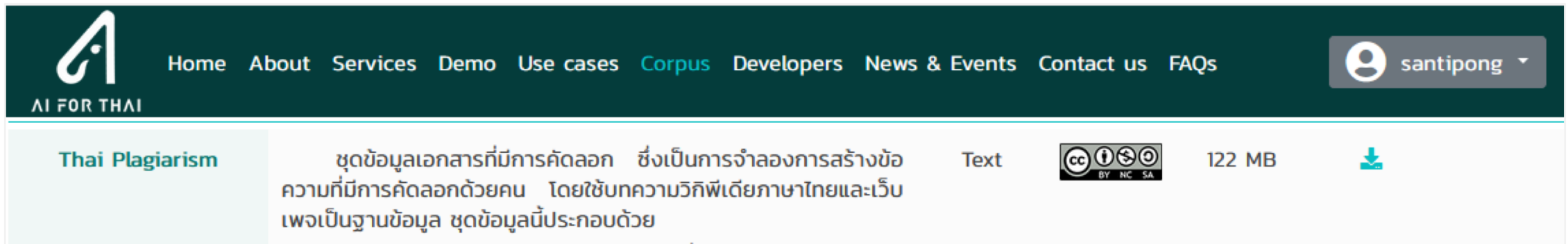
A computer consists of processing elements, typically a central processing unit, and some form of memory.

#### Modified document

A computer consists of processing elements, typically a central processing unit, and some form of memory (RAM).

# How to get the corpus?

- The corpus is free for research and development.
- You need to register to get the corpus.



The screenshot shows the AI FOR THAI website interface. The top navigation bar includes links for Home, About, Services, Demo, Use cases, Corpus (highlighted in teal), Developers, News & Events, Contact us, and FAQs. A user profile dropdown for 'santipong' is visible on the right. Below the navigation bar, the 'Thai Plagiarism' section is displayed, featuring a description in Thai: 'ชุดข้อมูลเอกสารที่มีการคัดลอก ซึ่งเป็นการจำลองการสร้างข้อความที่มีการคัดลอกด้วยคน โดยใช้บทความวิกิพีเดียภาษาไทยและเว็บเพจเป็นฐานข้อมูล ชุดข้อมูลนี้ประกอบด้วย'. To the right of the text, it indicates 'Text' format, a Creative Commons license (CC BY-NC-SA), and a file size of '122 MB'. A download icon is also present.

<https://aiforthai.in.th>



# Conclusions

- Designing and developing the tool that **imitates the plagiarism scenario** from plagiarists.
- Providing a Thai plagiarism annotation tool called **PlaTool** and a **Thai plagiarism guideline**.
- The corpus is manually created by **linguists**.
- The corpus is based on **four classes** of Thai plagiarism and linguistic mechanisms.

# Future work

- Rapidly increasing the size of corpus using the [artificial plagiarism method](#)
- Evaluating the performance of [plagiarism detection algorithms](#) in Thai
- Encouraging the Thai plagiarism corpus to be available as the [benchmark corpus](#)

# Thank you

## Kanokorn Trakultaweekoon

Speech and Text Understanding Research Team (STU)  
Artificial Intelligence Research Unit (AINRU)  
National Electronics and Computer Technology Center (NECTEC)

---



**Email:** [kanokorn.trakultaweekoon@nectec.or.th](mailto:kanokorn.trakultaweekoon@nectec.or.th)