วัตถุประสงค์ของการวิจัยครั้งนี้ เพื่อต้องหาสายอักขระเฉพาะสำหรับใช้ในการระบุภาษาของคำโดยใช้คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่นและภาษาฝรั่งเศส และพัฒนาระบบการระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างประเทศโดยใช้สายอักขระเฉพาะและใช้แบบจำลองเอ็นแกรมขนาด 1-5 แกรม

คลังข้อมูลที่ใช้ในงานวิจัยนี้ คือ คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่น ภาษาละ 10,000 คำ และคำทับศัพท์ภาษาฝรั่งเศส 1,000 คำ โดยเก็บจากข้อมูลที่พบในภาษาธรรมชาติซึ่งอาจจะไม่ได้ทับศัพท์ถูกต้องตามเกณฑ์ของราชบัณฑิตยสถานก็ได้ 80% ของคลังข้อมูลถูกนำมาใช้เพื่อหาสายอักขระเฉพาะและสร้างแบบจำลองเอ็นแกรมของแต่ละภาษา ในขณะที่อีก 20% ถูกใช้เพื่อการทดสอบระบบแบบต่างๆ

สายอักขระเฉพาะที่พบสะท้อนให้เห็นถึงลักษณะเฉพาะของแต่ละภาษาได้ในระดับหนึ่ง จึงมีผลให้ระบบที่ใช้สายอักขระเฉพาะในการระบุภาษาสามารถตัดสินภาษาได้ถูกต้อง 50.58% 48.71% 54.09% และ 20.40% สำหรับคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่น และฝรั่งเศส ตามลำดับ

เมื่อใช้แบบจำลองเอ็นแกรมในการระบุภาษา ระบบสามารถระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่นได้ถูกต้องกว่า 90% แต่ได้เพียงประมาณ 60% สำหรับคำทับศัพท์ฝรั่งเศส ผลที่ได้ยืนยันว่าขนาดของข้อมูลการฝึกมีผลต่อการทำงานของระบบการระบุภาษาทั้งสองระบบ นอกจากนี้ จากผลที่พบว่าระบบที่ใช้แบบจำลอง 3-แกรมให้ผลดีกว่าระบบที่ใช้ขนาดแกรมอื่นๆ ทำให้สรุปได้ว่า ขนาดของเอ็นแกรมมีผลต่อการทำงานของระบบการระบุภาษา

This research aims to find the unique character sequences of Thai and transliterated words (English, Japanese, and French), and implement language identification systems using unique character sequences and n-gram models (1-5 gram).

The corpora in this research consist of 10,000 Thai words, 10,000 English transliterated words, 10,000 Japanese transliterated words, and 1,000 French transliterated words. Transliterated words are collected from naturally occurring texts, even some of them are not conformed to the Royal Institute guidelines of transliteration. 80% of the corpus is used to extract unique character sequences and to build an n-gram language model of each language, while the other 20% is used for testing the systems.

The unique character sequences reflect some characteristics of the languages. As a result, the system using unique character sequence can identify languages correctly 50.58%, 48.71%, 54.09%, and 20.40% for Thai words, English, Japanese, and French transliterated words respectively.

When an n-gram language model is used, the system can identify languages correctly more than 90% for Thai, English and Japanese transliterated word, but only about 60% for French transliterated words. This confirms that the size of training corpus affects the performances of both systems. The results also show that the system using 3-gram model performs better than other n-gram models. Therefore, we can conclude that the size of n-gram does affect the performance of the language identification system.