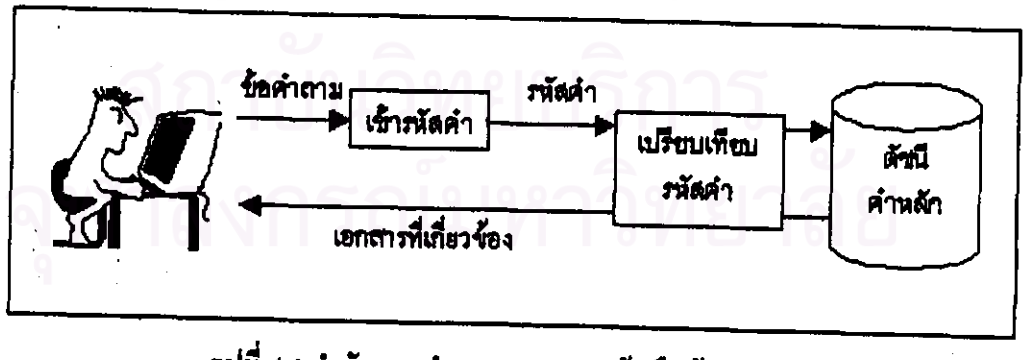
4. ขั้นตอนวิธีการค้นคืนข้ามภาษาแบบภาษาอังกฤษทับศัพท์ภาษาไทย

จากบทที่แล้วได้กล่าวถึงขั้นตอนวิธีสำหรับการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษ ในบทนี้จะกล่าวถึงการค้นคืนในส่วนภาษาอังกฤษทับศัพท์ภาษาไทย โดยจะแบ่งการทำงานของขั้นตอนวิธีออกเป็น 2 ขั้นตอน คือ ขั้นตอนวิธีการเข้ารหัสคำ และขั้นตอนวิธีการเปรียบเทียบรหัสคำ สำหรับการค้นคืน

4.1 โครงสร้างของระบบค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย

ขั้นตอนวิธีที่นำเสนอสำหรับการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษในส่วนภาษาอังกฤษทับศัพท์ภาษาไทยนั้นมีลำดับการทำงานดังนี้ เมื่อผู้ใช้งานได้ระบุข้อคำถามที่เป็นภาษาอังกฤษทับศัพท์ภาษาไทยให้กับระบบค้นคืนข้ามภาษาแล้ว ระบบจะทำการเข้ารหัสคำในข้อคำถาม เมื่อได้รหัสคำแล้วจะนำไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี คำหลักใดที่ผ่านเงื่อนไขการเปรียบเทียบจะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่งดังแสดงในรูปที่ 4.1 ขั้นตอนวิธีที่ได้นำเสนอนี้เปิดโอกาสให้ผู้ใช้สามารถระบุข้อคำถามที่เป็นภาษาอังกฤษทับศัพท์ภาษาไทยเพื่อค้นคืนเอกสารที่เป็นภาษาไทยและข้อคำถามที่เป็นภาษาไทยเพื่อค้นคืนเอกสารที่เป็นภาษาอังกฤษทับศัพท์ภาษาไทย โดยขั้นตอนวิธีนี้สามารถแบ่งออกเป็น 2 ส่วนคือ ขั้นตอนวิธีการเข้ารหัสคำ และขั้นตอนวิธีการเปรียบเทียบรหัสคำ



รูปที่ 4.1 ลำดับการทำงานของระบบค้นคืนข้ามภาษา

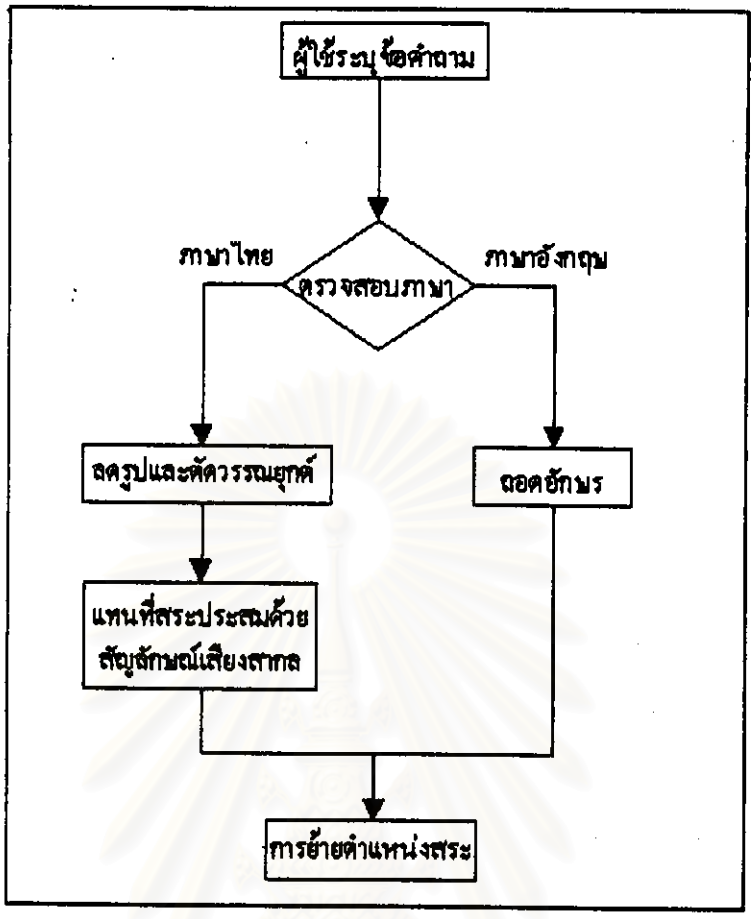
4.2 ขั้นตอนวิธีการเข้ารหัสคำ

ขั้นตอนวิธีการเข้ารหัสคำที่นำเสนอนั้นแบ่งออกเป็นสองขั้นตอนย่อย คือ ขั้นตอนการประมวลผลตัวอักษรเบื้องต้นและการย้ายตำแหน่งสระ เพื่อให้สอดคล้องกับโครงสร้างของระบบค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทยที่กล่าวไว้ในหัวข้อ 4.1 ซึ่งขั้นตอนการประมวลผลตัวอักษรเบื้องต้นจะตรวจสอบข้อคำถามที่ผู้ใช้ระบุมาว่าเป็นภาษาอังกฤษทับศัพท์ภาษาไทยหรือเป็นภาษาไทยโดยจะเปลี่ยนให้ข้อคำถามทั้งสองแบบนี้อยู่ในรูปแบบเดียวกันก่อนจึงจะทำการย้ายตำแหน่งสระต่อไป

4.2.1 การประมวลผลตัวอักษรเบื้องต้น

เพื่อให้การเข้ารหัสคำไทยหรือคำอังกฤษทับศัพท์ได้รหัสตรงกันนั้น งานวิจัยนี้ได้อาศัยหลักการอ่านออกเสียงที่เหมือนกัน ซึ่งการอาศัยแต่หน่วยอักษรบางครั้งไม่สามารถบอกได้ว่ามีหน่วยเสียงที่ถูกต้องเป็นอย่างไร เช่น หน่วยอักษร “ร” จะถอดเป็นหน่วยเสียง /ร/ แต่ถ้ามี “ร” สองตัวติดกันอาจจะเป็น ร หัน ซึ่งมีหน่วยเสียงเป็น / ั / ในกรณีที่มีตัวสะกดตามหลัง หรือมีหน่วยเสียงเป็น / ัน / ในกรณีที่ไม่มีตัวสะกดตามหลัง เป็นต้น ดังนั้นผู้วิจัยจึงได้นำเสนอขั้นตอนการประมวลผลตัวอักษรเบื้องต้นเพื่อแก้ปัญหาในเรื่องของการอ่านออกเสียงของคำในภาษาไทย และทำให้ข้อคำถามที่เป็นภาษาไทยหรือภาษาอังกฤษทับศัพท์ภาษาไทยอยู่ในรูปแบบเดียวกัน เพื่อให้ง่ายต่อการประมวลผลต่อไป

การประมวลผลตัวอักษรเบื้องต้น ประกอบด้วย การถอดรูปและตัดวรรณยุกต์ การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล และการถอดอักษร โดยจะพิจารณาจากข้อคำถามที่ผู้ใช้ระบุให้กับระบบ ถ้าข้อคำถามเป็นภาษาไทยจะต้องทำการการประมวลผลตัวอักษรเบื้องต้นในส่วนของการถอดรูปและตัดวรรณยุกต์ และการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล แต่ถ้าข้อคำถามเป็นภาษาอังกฤษทับศัพท์ภาษาไทยจะต้องทำการประมวลผลตัวอักษรเบื้องต้นในส่วนของการถอดอักษร ก่อนดังแสดงในรูปที่ 4.2



รูปที่ 4.2 การประมวลผลตัวอักษรเบื้องต้น

4.2.1.1 การถอดรูปและตัดวรรณยุกต์

การถอดรูปและตัดวรรณยุกต์เป็นการประมวลผลตัวอักษรเบื้องต้นส่วนหนึ่งที่พยายามเปลี่ยนรูปแบบการสะกดของคำไทยให้อยู่ในรูปแบบที่ง่ายต่อการประมวลผลในขั้นต่อไป โดยพยายามทำให้ผลลัพธ์ที่ได้จากขั้นตอนนี้มีรูปแบบเหมือนกันผลลัพธ์ที่ได้จากการถอดอักษรคำภาษาอังกฤษทับศัพท์ภาษาไทย ผู้วิจัยได้นำเสนอขั้นตอนวิธีการอ่านออกเสียงคำไทยโดยแบ่งเป็นกรณีต่าง ๆ ดังนี้

- <u>วรรณยุกต์ และไม้ไต่คู้</u> ทำการตัดวรรณยุกต์และไม้ไต่คู้ทิ้ง เนื่องจากในการถอดอักษรไทยเป็นอักษรอังกฤษ จะไม่พิจารณาวรรณยุกต์และไม้ไต่คู้ เช่น ช้าง ถอดอักษรเป็น CHANG
- <u>ร ร</u> ทำการเปลี่ยน รร เป็น ั น ในกรณีที่ไม่มีตัวสะกดตามหลัง เช่น จรรยา บรรจบ ครรภ์ เป็นต้น และเป็น ั ร ในกรณีที่มีตัวสะกดตามหลัง เช่น ธรรม พรรณ กรรม

เป็นต้น จากการวิเคราะห์รูปแบบของคำผู้วิจัยได้สร้างรูปแบบของ รร โดยยึดหลักทางภาษาศาสตร์[1] และข้อมูลทางสถิติ[2] ซึ่งมีรูปแบบดังนี้

กำหนดให้

... อักขระใด	<V> สระ
<E> จุดสิ้นสุดสายอักขระ	<IV> สระหน้า คือ เ- แ- ไ- ใ- และโ-
<C> พยัญชนะ	

รูปแบบ รร ที่เปลี่ยนเป็น ั-น

...รร<E>	เช่น สรร
...รร<C><C><E>	เช่น บรรทม
...รร<C>์...	เช่น สรรค์
...รร<IV>...	เช่น บรรเลง
...รร<C><V>...	เช่น สรรชัย

รูปแบบ รร ที่เปลี่ยนเป็น ั

...รร<C><E>	เช่น ธรรม
...รร<C><IV>...	เช่น บรรณโลก

- <u>สระ ใ- ไ- และไ-ย</u> ทำการเปลี่ยนสระดังกล่าวทั้งหมดเป็น ั-ย รูปแบบเดียวเพื่อเพิ่มความถูกต้องยิ่งขึ้นในการเข้ารหัสคำ เนื่องจากสระดังกล่าวมีการอ่านออกเสียงที่คล้ายกันมาก และในการถอดอักษรมักจะถอดเป็น –AI เหมือนกัน เช่น ใช ไช ไชย และชัย ต่างก็ถอดอักษรเป็น CHAI เป็นต้น การเปลี่ยนสระ ใ- ไ- และไ-ย มีหลักเกณฑ์ต่าง ๆ ดังนี้

  1. การเปลี่ยนสระ ใ- เป็น ั-ย เนื่องจากสระ ใ- ที่ใช้กับอักษรควบซึ่งมีทั้งหมด 5 คำ[3] คือ ใกล้ ใหม่ ใหล ใหญ่ และใคร(ใคร่) โดยจะเปลี่ยน ใ เป็น ั-ย หลัง

---

[1] กำชัย ทองหล่อ, <u>หลักภาษาไทย</u>, พิมพ์ครั้งที่ 10 (กรุงเทพมหานคร : อมรการพิมพ์, 2540).

[2] S. Sethaputra, <u>Thaisoft So Sethaputra Dictionary Version 1.5</u> [Computer Software], Bangkok : Thaisoft Co.,Ltd, 1996.

[3] กำชัย ทองหล่อ, <u>หลักภาษาไทย</u>, พิมพ์ครั้งที่ 10 (กรุงเทพมหานคร : อมรการพิมพ์, 2540).

พยัญชนะควบ (หลังจากทำการตัดวรรณยุกต์แล้ว) เช่น ใกล้ ใหม่ และใหญ่ จะเปลี่ยนเป็น กลัย หมัย และหญัย ตามลำดับ นอกเหนือจาก 5 คำนี้แล้วจะเปลี่ยน ไ เป็น -ัย หลังพยัญชนะต้นของสระ ไ- เช่น ใน ใจ และใส จะเปลี่ยนเป็น นัย จัย และสัย ตามลำดับ

2. การเปลี่ยนสระ ไ-ย เป็น -ัย เนื่องจากสระ ไ-ย ที่มีใช้ส่วนใหญ่เป็นคำที่มาจากบาลี-สันสกฤต ซึ่งมีใช้อยู่ 7 คำ [4] คือ ไชย ไทย ไตย พิมพิลาไลย ภาคิไนย อภิไชย อสงไขย โดยจะเปลี่ยน ไ- เป็น -ั หลังพยัญชนะต้นส่วนอักษร ย มีอยู่แล้ว
3. การเปลี่ยนสระ ไ- ที่ใช้กับอักษรควบจะเปลี่ยน ไ เป็น -ัย หลังพยัญชนะควบ (หลังจากทำการตัดวรรณยุกต์แล้ว) เช่น ไกร ไตร และไทร จะเปลี่ยนเป็น กรัย ตรัย และทรัย ตามลำดับ ซึ่งรูปแบบที่กำหนดนี้

...ไ<C><C><E> เช่น ไกร
...ไ<C><C><IV>... เช่น ไกล่เกลี่ย

นอกจากนี้แล้วจะถือว่าเป็นสระ ไ- ที่ใช้กับพยัญชนะธรรมดาซึ่งจะเปลี่ยนเป็น -ัย หลังพยัญชนะต้น

- สระ -ำ และ -ัม ทำการเปลี่ยนสระ -ำ เป็น -ัม รูปแบบเดียวเพื่อเพิ่มความถูกต้องยิ่งขึ้นในการเข้ารหัสคำ เนื่องจากสระดังกล่าวมีการอ่านออกเสียงที่เหมือนกันและในการถอดอักษรมักจะถอดเป็น –AM เหมือนกัน เช่น กำ และ กัม ต่างก็ถอดอักษรเป็น KAM เป็นต้น

- การันต์และอักษรควบการันต์ ทำการตัดอักษรและอักษรควบที่มีตัวการันต์กำกับออก เนื่องจากการถอดอักษรโดยปกติจะไม่ถอดอักษรที่มีการันต์กำกับ เช่น สิทธิ์ ถอดเป็น SITH พันธุ์ ถอดเป็น PHAN เป็นต้น การตัดอักษรและอักษรควบการันต์ออกมี 7 กรณี [5] คือ

---

[4] S. Sethaputra, Thaisoft So Sethaputra Dictionary Version 1.5 [Computer Software], Bangkok : Thaisoft Co.,Ltd, 1996.

[5] กำชัย ทองหล่อ, หลักภาษาไทย, พิมพ์ครั้งที่ 10 (กรุงเทพมหานคร : อมรการพิมพ์, 2540).

1. สระเสียงสั้นที่ไม่มีตัวสะกดควบการันต์ คือ พยัญชนะตามด้วย สระ อิ หรือ สระ อุ และการันต์ โดยจะตัดตัวสะกด สระ และการันต์ทิ้ง ตัวอย่างเช่น คำว่า สิทธิ์ และพันธุ์ จะตัดเป็น สิท และ พัน ตามลำดับ
2. พยัญชนะตัวเดียว คือ มีพยัญชนะหนึ่งตัวและการันต์ โดยจะตัดพยัญชนะหนึ่งตัวและการันต์ ตัวอย่างเช่น คำว่า กรณ์ สัตย์ ยุทธ์ และทรัพย์ จะถูกตัดเป็น กร สัต ยุท และทรัพ ตามลำดับ
3. พยัญชนะสองตัว คือ -จน์ โดยจะตัด จน์ ทิ้ง ตัวอย่างเช่น คำว่า กาญจน์ สิญจน์ และวัญจน์ จะตัดเป็น กาญ สิญ และ วัญ ตามลำดับ
4. พยัญชนะสามตัว คือ –ษมณ์ โดยจะตัด ษมณ์ ทิ้ง ตัวอย่างเช่น ลักษมณ์ จะตัดเป็น ลัก
5. อักษรควบแท้ จะตัดอักษรควบทั้งสองตัวและการันต์ ตัวอย่างเช่น พักตร์ และ อินทร์ จะตัดเป็น พัก และอิน ตามลำดับ
6. อักษรควบไม่แท้ จะตัดพยัญชนะหนึ่งตัวและการันต์ ตัวอย่างเช่น คำว่า ธรรม์ สรรพ์ และหรรษ์ จะถูกตัดเป็น ธรร สรร และหรร ตามลำดับ
7. อักษรนำ จะตัดพยัญชนะทั้งสองตัวและการันต์ ตัวอย่างเช่น คำว่า ลักขณ์ จะถูกตัดเป็น ลัก

- <u>สระ ฤ และ ฤๅ</u> ทำการเปลี่ยน ฤ เป็น รึ และ ฤๅ เป็น รือ เช่น พฤกษา เปลี่ยนเป็น พรึกษา ฤาษี เปลี่ยนเป็น รือษี เป็นต้น ส่วน ฦ และ ฦๅ ไม่นำมาพิจารณา เนื่องจากปัจจุบันไม่มีที่ใช้แล้ว [6]

<u>หมายเหตุ</u> ทุกกรณีที่กล่าวมานั้นจะต้องทำงานตามลำดับก่อนหลัง เพื่อลดการตรวจสอบซ้ำซ้อน เช่น จะทำการวรรณยุกต์ และไม้ไต่คู้ก่อน และในส่วนของ รร จะไม่พิจารณาว่ามีวรรณยุกต์อยู่ เป็นต้น

หลังจากที่นำคำไทยผ่านขั้นตอนวิธีการลดรูปและตัดวรรณยุกต์เพื่อเปลี่ยนคำไทยให้อยู่ในรูปแบบที่ง่ายต่อการประมวลผลแล้ว จะทำการแทนที่สระสระประสมด้วยสัญลักษณ์เสียงสากล ซึ่งจะกล่าวในหัวข้อถัดไป

---

[6] กำชัย ทองหล่อ. <u>หลักภาษาไทย</u>. พิมพ์ครั้งที่ 10 (กรุงเทพมหานคร : อมรการพิมพ์, 2540).

4.2.1.2 การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล

การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากลเป็นการประมวลผลตัวอักษรเบื้องต้นส่วนหนึ่ง เพื่อทำให้คำในภาษาไทยที่มีการใช้สระประสมและสระเดี่ยวที่ใช้อักขระตั้งแต่สองตัวขึ้นไป อยู่รูปแบบที่ง่ายต่อการประมวลผล และลดจำนวนอักขระที่ต้องเปรียบเทียบ ซึ่งขั้นตอนนี้จะใช้สัญลักษณ์เสียงสากลแทนเสียงสระดังกล่าวโดยจะวางสัญลักษณ์เสียงสากลไว้หลังพยัญชนะต้นเพื่อให้มีรูปแบบเหมือนกับผลลัพธ์ที่ได้จากการถอดอักษรคำ การใช้สัญลักษณ์เสียงสากลแทนที่สระมีดังนี้

- ใช้สัญลักษณ์เสียง e แทนสระ เ-ะ เช่น เละ เปลี่ยนเป็น ne
- ใช้สัญลักษณ์เสียง x แทนสระ แ-ะ แ--ะ เช่น แกะ แหละ เปลี่ยนเป็น nx หnx ตามลำดับ
- ใช้สัญลักษณ์เสียง q แทนสระ เ- เ-ิ เช่น เกิด เผชิญ เปลี่ยนเป็น nqd ผชqญ ตามลำดับ
- ใช้สัญลักษณ์เสียง I (ดัดแปลงจาก ia) แทนสระ เ-ียะ เ-ียะ เ-ีย และ เ-ีย เช่น เสียง เกรียน เปลี่ยนเป็น sIง กวIน ตามลำดับ
- ใช้สัญลักษณ์เสียง U (ดัดแปลงจาก ua) แทนสระ เ-ือะ เ-ือ เ-ือ ัวะ และ ัว เช่น เรือง เกลือ ผัวะ และ วัว เปลี่ยนเป็น รUง กลU ผU และ วU ตามลำดับ
- ใช้สัญลักษณ์เสียง @ แทนสระ เ-า เ-า เ-าะ และ เ-าะ เช่น เงา เพรา เงาะ และ เฉพาะ เปลี่ยนเป็น ง@ พร@ ง@ และ ฉพ@ ตามลำดับ

หมายเหตุ สามารถดูรายละเอียดการใช้สัญลักษณ์เสียงสากลสำหรับอักขระไทยได้ในภาคผนวก ข

4.2.1.3 การถอดอักษร

ในกรณีที่ข้อคำถามเป็นภาษาอังกฤษทับศัพท์ภาษาไทยจะทำการถอดอักษรอังกฤษเป็นภาษาไทย การถอดอักษรที่นำเสนอในที่นี้ จะเป็นการเปลี่ยนพยัญชนะอังกฤษเป็นพยัญชนะไทย ส่วนสระอังกฤษจะใช้หลักเกณฑ์เหมือนกับการตรวจหาสระประสม คือจะถอดสระอังกฤษเป็นสระไทย แต่ถ้าสระไทยนั้นเป็นสระประสมหรือสระเดี่ยวที่ใช้อักขระตั้งแต่สองตัวขึ้นไปจะใช้สัญลักษณ์เสียงสากล (หนึ่งตัว) แทนเสียงสระดังกล่าว

หลักเกณฑ์ในการถอดในส่วนของพยัญชนะจะใช้หลักการเทียบตัวอักษรโรมัน-ไทย ของ ISO [7] ดังตารางที่ 4.1 ยกเว้นอักษรบางตัวที่ ISO ไม่ได้เทียบให้ ได้แก่อักษร G Q และ X หรือบางตัวที่เทียบให้แล้วแต่ไม่นิยมใช้ เช่น จ เท่ากับ C ซึ่งส่วนใหญ่นิยมใช้ CH หรือ J สำหรับ จ [8] ดังนั้นผู้วิจัยจึงนำเสนอการถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของพยัญชนะ ดังที่แสดงในตารางที่ 4.2 ซึ่งได้ทำการปรับเปลี่ยนและเพิ่มเติมบางตัวจากต้นแบบของ ISO (ที่มีเครื่องหมาย * ในตาราง) เพื่อให้สมบูรณ์ยิ่งขึ้น เช่น ถอด DH เป็น ท และ ถอด BH เป็น พ เนื่องจากเป็นการเขียนแบบบาลีสันสกฤต ซึ่งยังมีใช้กันอยู่มากในปัจจุบัน [9]

---

[7] International Organization for Standardization, "Information and Documentation – Transliteration of Thai", Draft International Standard ISO/DIS 11940, 1996.

[8] จันทร์เพ็ญ ไวหารสุนทร, "การศึกษาการใช้อักษรโรมันแทนอักษรไทย" (ปริญญานิพนธ์ คณะอักษรศาสตร์ มหาวิทยาลัยศรีนครินทร์วิโรฒ ประสานมิตร, 2530).

[9] พจนต์ ทิมเจริญ, "การเขียนชื่อภาษาไทยด้วยอักษรโรมัน," วารสารแผนที่ ปีที่ 27 ฉบับที่ 2 (ตุลาคม-ธันวาคม 2527) : 61-74.

พยัญชนะไทย	อักษรโรมัน
ก	K
ข	KH
ฃ	KH
ค	KH
ฅ	KH
ฆ	KH
ง	NG
จ	C
ฉ	CH
ช	CH
ซ	S
ฌ	CH
ญ	Y
ฎ	D
ฏ	T
ฐ	TH
ฑ	TH
ฒ	TH
ณ	N
ด	D
ต	T
ถ	TH
ท	TH

พยัญชนะไทย	อักษรโรมัน
ธ	TH
น	N
บ	B
ป	P
ผ	PH
ฝ	F
พ	PH
ฟ	F
ภ	PH
ม	M
ย	Y
ร	R
ฤ	R
ล	L
ฦ	L
ว	W
ศ	S
ษ	S
ส	S
ห	H
ฬ	L
อ	X
ฮ	H

ตารางที่ 4.1 การใช้อักษรโรมันแทนพยัญชนะไทยของ ISO

อักษรอังกฤษ	อักษรไทย	หมายเหตุ
B	บ	
BH	พ	*
C	ช	*
CH	ช (ฉ ฌ)	
CK	ก	*
D	ด (ฎ)	
DH	ท	*
F	ฟ่ (ฝ)	
G	ก	*
H	ห (ฮ)	
J	จ	*
K	ก	
KH	ข (ฃ ค ฅ ฆ)	
L	ล (ฦ ฬ)	
M	ม	

อักษรอังกฤษ	อักษรไทย	หมายเหตุ
N	น (ณ)	
NG	ง	
P	ป	
PH	พ (ผ ภ)	
Q	ค	*
R	ร (ฤ)	
S	ส (ซ ศ ษ)	
T	ต (ฏ)	
TH	ท (ฐ ฑ ฒ ถ ธ)	
V	ว	
W	ว	
X	ก	*
Y	ย (ญ)	
Z	ซ	*

* ส่วนที่ปรับเปลี่ยนและเพิ่มเติมจากแบบของ ISO

ตารางที่ 4.2 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของพยัญชนะที่นำเสนอ

ส่วนหลักเกณฑ์ในการถอดอักษรในส่วนสระอังกฤษเป็นสระไทยนั้น ผู้วิจัยพบว่ามีปัญหาอย่างมากในการถอดอักษรคือ หนึ่งหน่วยอักษรโรมันสามารถถอดได้เป็นหลายหน่วยอักษรไทย เช่น A ถอดอักษรเป็น เ- แ- -ะ -า และ O ถอดอักษรเป็น -อ โ- เป็นต้น และหลายหน่วยอักษรโรมันสามารถถอดเป็นหนึ่งหน่วยอักษรไทย เช่น U หรือ OO ถอดเป็น -ุ เป็นต้น และจากการศึกษาของผู้วิจัยพบว่ามีความหลากหลายในการใช้อักษรโรมันแทนอักษรไทย เช่น คำว่า พร มีการเขียนเป็น Phorn Phon Porn Pon เป็นต้น ซึ่งการใช้ orn เป็นความนิยมใช้ซึ่งไม่ถูกต้องตามหลักภาษาที่ราชบัณฑิตสถานกำหนด ดังนั้นในงานวิจัยนี้จึงได้พยายามยึดหลักการถอดอักษรของราชบัณฑิตยสถาน [10] และ การใช้อักษรโรมันแทนอักษรไทยของจันทร์เพ็ญ โวหารสุนทร[11] (ได้จาก

[10] ราชบัณฑิตยสถาน, "ประกาศราชบัณฑิตยสถาน เรื่อง การถอดอักษรไทยเป็นโรมัน", 2482.

[11] จันทร์เพ็ญ โวหารสุนทร, "การศึกษาการใช้อักษรโรมันแทนอักษรไทย" (ปริญญานิพนธ์ คณะอักษรศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร, 2530).

การสำรวจความนิยมในการใช้อักษรโรมันแทนอักษรไทย) เป็นต้นแบบและเพิ่มเติมบางส่วนจากการศึกษาของผู้วิจัยที่มีผู้นิยมใช้เข้าไป (ที่มีเครื่องหมาย * ในตาราง) เพื่อความถูกต้องในการใช้งาน ดังที่แสดงในตารางที่ 4.3

ตัวอักษร	ถอดอักษรเป็น	หมายเหตุ
-A	= ะ	
-AA	= า	*
-AE	= x (เ-ะ แ-)	
-AI	= ัย	
-AO	= @ (เ-า)	
-AIU	= I (เ-ียว)	*
-ARN	= าน	*
-ART	= าท	*
-E	= เ (เ-ะ เ-)	
-EE	= ี	*
-EO	= แ-ว	
-ER	= q (เ-อ เ-ิ) หลัง R ต้องไม่เป็นสระ	
-EU	= ึ	
-I	= ิ	
-IA	= I (เ-ียะ เ-ีย)	
-IE	= I (เ-ียะ เ-ีย)	*
-O	= อ (-อ)	
-OE	= q (เ-อ เ-ิ)	
-OI	= อย	
-OO	= ู	
-ORN	= ร (-อน)	*
-U	= ุ (ึ ึ ุ ู)	
-UA	= U (เ-ือะ เ-ือ ัวะ ัว)	
-UE	= ึ	

* ส่วนที่เพิ่มเติมจากแบบของราชบัณฑิตยสถาน

ตารางที่ 4.3 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของสระที่นำเสนอ

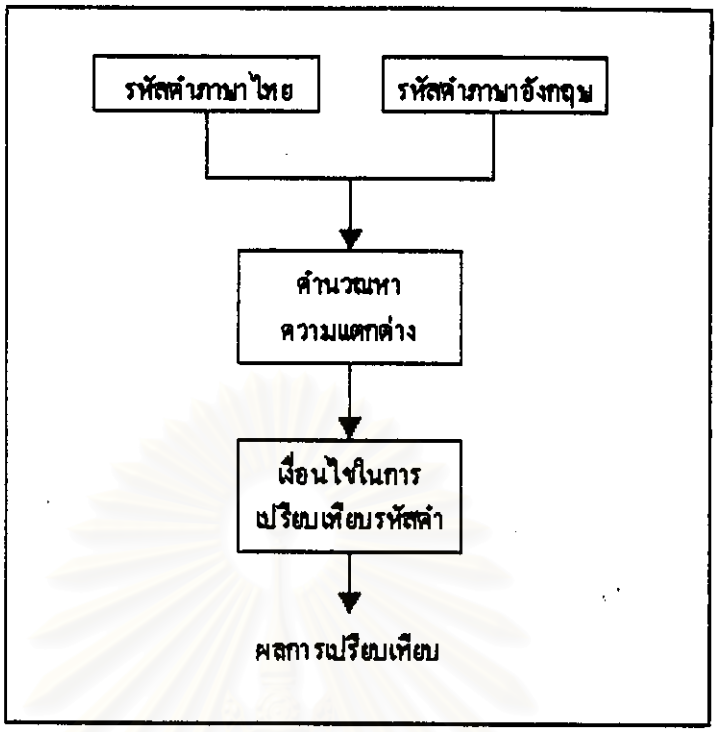
<u>หมายเหตุ</u> ถ้าตัวอักษรแรกของคำเป็นสระได้แก่ A E I O และ U ให้ถอด A เป็น อ ถอด E เป็น เอ (สระ เ- กับ อ) ถอด I เป็น อิ (อ กับสระ ิ) ถอด O เป็น โอ (สระ โ- กับ อ) และถอด U เป็น อุ (อ กับสระ ุ) และถ้าตัวอักษรถัดไปเป็นสระอีกให้นำอักษรตัวแรกไปรวมด้วยในการถอดอักษรโดยเทียบตามตารางที่ 4.3

4.2.2 การย้ายตำแหน่งสระ

หลังจากคำภาษาอังกฤษทับศัพท์กาษาไทยและคำภาษาไทยที่ได้ผ่านขั้นตอนประมวลผลตัวอักษรเบื้องต้นแล้ว คำศัพท์ที่ได้จะอยู่ในรูปแบบเดียวกันคือ อักษรไทยผสมกับสัญลักษณ์เสียงสากล (แทนเสียงสระ) ขั้นตอนสุดท้ายของการเข้ารหัสคำคือ การย้ายตำแหน่งสระ โดยจะทำการย้ายสระไปข้างหลังของสายอักขระตามลำดับ เช่น อุดมศึก เข้ารหัสเป็น อดมศก ุ ึ เป็นต้น

4.3 ขั้นตอนวิธีเปรียบเทียบรหัสคำ

ขั้นตอนวิธีในการเปรียบเทียบรหัสคำที่นำเสนอในที่นี้มีขั้นตอนดังนี้ เริ่มจากนำรหัสคำทั้งสองภาษามาทำการคำนวณหาค่าความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด (Minimum Edit Distance) และนำค่าความแตกต่างของคู่รหัสคำ มาทดสอบกับเงื่อนไขในการเปรียบเทียบ ถ้าผ่านการทดสอบจะถือว่ารหัสคำทั้งสองรหัสนี้เป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา ดังรูปที่ 4.3



รูปที่ 4.3 การเปรียบเทียบรหัสคำที่นำเสนอ

การคำนวณหาค่าความแตกต่าง จะพิจารณาความแตกต่างกันที่เสียงของอักขระไม่ใช่ความแตกต่างที่รูปของอักขระ ส่วนเงื่อนไขในการเปรียบเทียบรหัสคำนั้นจะนำเอาความยาวของรหัสคำมาร่วมพิจารณากับค่าความแตกต่างของรหัสคำด้วย ซึ่งรายละเอียดของขั้นตอนนั้นจะกล่าวในหัวข้อถัดไป

4.3.1 การคำนวณหาค่าความแตกต่างของรหัสคำ

ผู้วิจัยได้นำเสนอการคำนวณหาค่าความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขเสียงอ่านสั้นที่สุด (Mininum Phonetic Edit Distance) ของ J. Zobel [12] ซึ่งมีขั้นตอนวิธีการคำนวณคล้ายกับเทคนิคระยะแก้ไขสั้นที่สุด คือการคำนวณหาค่าความแตกต่างของคำจะได้มาจากการคำนวณหาต้นทุนน้อยที่สุดในการแก้ไขอักขระให้คำทั้งสองเหมือนกัน แต่เทคนิคระยะแก้ไขเสียงอ่านสั้นที่สุดจะพิจารณาความแตกต่างกันทางเสียงของอักขระแทนรูปของอักขระ โดยใช้กลุ่มอักขระของ

---

[12] J. Zobel and P. Dart, Phonetic String Matching: Lessons from Information Retrieval, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 166-172, 1996.

ซาวด์เด็กซ์ช่วยในการกำหนดกลุ่มเสียงที่คล้ายกัน การคำนวณค่าความแตกต่างสามารถแสดงในรูปของสมการเวียนเกิด Edit (P_j, W_k) ได้ดังนี้

$$
\begin{aligned}
Edit(P_0, W_0) &= 0 \\
Edit(P_j, W_0) &= Edit(P_{j-1}, W_0) + D(p_{j-1}, p_j) \\
Edit(P_0, W_k) &= Edit(P_0, W_{k-1}) + D(w_{k-1}, w_k) \\
Edit(P_j, W_k) &= \min[\ Edit(P_{j-1}, W_k) + D(p_{j-1}, w_j), \\
&\qquad Edit(P_j, W_{k-1}) + D(p_{k-1}, w_k), \\
&\qquad Edit(P_{j-1}, W_{k-1}) + R(p_j, w_k)\ ]
\end{aligned}
$$

โดยที่

$P_j = p_1 p_2 p_3 \ldots p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1 w_2 w_3 \ldots w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$R(p_j, w_k)$ เป็นฟังก์ชันที่กำหนดต้นทุนในการแทนที่อักขระ โดยแบ่งออกเป็น 3 กรณี คือ (1) ต้นทุนเท่ากับ 0 ถ้า p_j และ w_k เป็นอักขระที่เหมือนกัน (2) ต้นทุนเท่ากับ 1 ถ้า p_j และ w_k เป็นอักขระที่อยู่กลุ่มเสียงเดียวกันในซาวด์เด็กซ์ (ดูรายละเอียดได้ในตารางที่ 2.2 การกำหนดรหัสซาวด์เด็กซ์) และ (3) นอกจากนี้แล้ว ต้นทุนจะเท่ากับ 2

$D(p_j, w_k)$ เป็นฟังก์ชันที่กำหนดต้นทุนในการเพิ่มหรือถบอักขระ โดยแบ่งออกเป็น 3 กรณี คือ (1) ต้นทุนเท่ากับ 0 ถ้า p_j และ w_k เป็นอักขระที่เหมือนกัน (2) ต้นทุนเท่ากับ 1 ถ้า p_j และ w_k เป็นอักขระที่อยู่กลุ่มที่ไม่ออกเสียง ได้แก่ อักขระ H และ W หรือเป็นอักขระที่อยู่กลุ่มเสียงเดียวกันในซาวด์เด็กซ์ (ตารางที่ 2.2 การกำหนดรหัสซาวด์เด็กซ์) และ (3) นอกจากนี้แล้ว ต้นทุนจะเท่ากับ 2

จากการศึกษาของผู้วิจัยพบว่าระยะแก้ไขเสียงอ่านสั้นที่สุดของ J. Zobel ยังขาดความเหมาะสมเมื่อนำมาใช้กับภาษาไทย ผู้วิจัยจึงได้ทำการปรับเปลี่ยนฟังก์ชันการกำหนดต้นทุนในการแก้ไขอักขระเพื่อความเหมาะสมเมื่อใช้กับคำในภาษาไทยโดยมีแนวคิดดังนี้

- แบ่งการคำนวณหาค่าความแตกต่างของรหัสคำออกเป็น 2 ส่วนคือ (1) ค่าความแตกต่างของรหัสคำในส่วนที่เป็นพยัญชนะและ (2) ค่าความแตกต่างของรหัสคำกับในส่วนที่เป็นสระ เนื่องจากในการถอดอักษรด้วยกฎที่ใช้ในงานวิจัยนี้พบว่าตำแหน่ง

ของสระที่ได้จากการถอดมักจะคลาดเคลื่อนไปจากตำแหน่งที่ควรจะเป็น ซึ่งจะทำให้การคำนวณหาค่าความแตกต่างของรหัสคำผิดพลาดด้วย

- กำหนดต้นทุนต่อการแก้ไขอักขระให้มี 4 ระดับ คือ C_1 C_2 C_3 และ C_4 โดยที่ C_1 จะหมายถึงการใช้ต้นทุนน้อยที่สุดในการแก้ไขอักขระ และ C_4 จะหมายถึงการใช้ต้นทุนมากที่สุดในการแก้ไขอักขระตามลำดับ ในที่นี้จะกำหนดให้ C_1=0 C_2=1 C_3=4 และ C_4=7 ซึ่งจะเพิ่มความละเอียดในการคำนวณหาค่าความแตกต่างของรหัสคำ
- กำหนดให้ต้นทุนในการแก้ไขแต่ละอักขระไม่เท่ากัน จากเดิมกำหนดให้การเพิ่ม การลบ และการแทนที่อักขระใด ๆ จะกำหนดต้นทุนในการแก้ไขเท่ากับหนึ่งหรือสองแล้วแต่กรณี ในขณะที่วิธีการที่นำเสนอจะกำหนดต้นทุนในการแก้ไขโดยจะพิจารณาจากความสำคัญของอักขระที่แก้ไข เช่น ต้นทุนในการลบสระ –ะ ควรจะน้อยกว่าต้นทุนการลบพยัญชนะ เนื่องจากภาษาไทยมีการลดรูปสระ –ะ แต่ภาษาอังกฤษทับศัพท์ภาษาไทยไม่มีการลดรูปเสียงสระ -ะ ต้นทุนในการแทนที่อักขระที่อยู่ในมาตราเดียวกันควรจะน้อยกว่าต้นทุนการแทนที่อักขระที่อยู่ต่างมาตรา เป็นต้น
- อนุญาตให้มีการแทนที่อักขระแบบหนึ่งอักขระต่อสองอักขระ จากเดิมหนึ่งอักขระต่อหนึ่งอักขระ เนื่องจากในภาษาไทยมีลักษณะการใช้อักษรควบ โดยอักษรควบที่ส่งผลต่อการอ่านออกเสียง เช่น การเปลี่ยน “จัก” ให้เป็น “จักร” ต้นทุนในการแทนที่ ก ด้วย กร ควรจะน้อยกว่าต้นทุนในการเพิ่ม ร

รายละเอียดต่าง ๆ ในการออกแบบฟังก์ชันกำหนดต้นทุนในการแก้ไขอักขระจะกล่าวในหัวข้อถัดไป

4.3.1.1 การกำหนดต้นทุนในการแก้ไขอักขระ

จากการศึกษาของผู้วิจัยพบว่าการใช้กลุ่มเสียงภาษาไทยของซาวด์เด็กซ์ร่วมกับการกำหนดต้นทุนในการแก้ไขอักขระ ทำให้การคำนวณค่าความแตกต่างคลาดเคลื่อนไปจากความเป็นจริงมาก เช่น ไม่มีการจัดการเรื่องอักษรควบ

ผู้วิจัยได้ออกแบบตารางการกำหนดต้นทุนในการแทนที่อักขระโดยกำหนดให้แต่ละอักขระมีต้นทุนที่แตกต่างกันแล้วแต่ตัวที่จะไปแทนที่แบบอักขระต่ออักขระ โดยมีแนวคิดดังนี้

- พยัญชนะไทยมี 44 รูป สามารถแยกความแตกต่างทางเสียงได้ 21 เสียง ดังนั้น พยัญชนะที่มีเสียงเหมือนกันตามหลักภาษาไทยจะให้ต้นทุนในการแทนที่อักขระเท่ากับ C_1 ดังตาราง ที่ 4.4

ก	ฎ ด	ฝ ฟ
ข ฃ ค ฅ ฆ	ฏ ต	ม
ง	ฐ ฑ ฒ ถ ท ธ	ร
จ	ณ น	ล ฬ
ฉ ช ฌ	บ	ว
ซ ศ ษ ส	ป	ห ฮ
ญ ย	ผ พ ภ	อ

ตารางที่ 4.4 กลุ่มเสียงของพยัญชนะไทย

- อักขระที่ไม่สามารถแยกความแตกต่างในการถอดอักษร โดยอักษรไทยเหล่านี้ไม่ได้อยู่กลุ่มเสียงเดียวกันตามตารางที่ 4.4 ซึ่งจะกำหนดต้นทุนในการแทนที่อักขระเหล่านี้เท่ากับ C_2 ดังตารางที่ 4.5

อักษรโรมัน	อักษรไทย
K	ก ค
T	ต ท
P	ป พ
CH	จ ช
A	-ะ -า
E	เ- -ี
U	-ุ -ื

ตารางที่ 4.5 อักขระที่ไม่สามารถแยกความแตกต่างในการถอดอักษร

- อักขระที่ไม่สามารถแยกความแตกต่างทางเสียงในภาษาอังกฤษ เนื่องจากการใช้อักษรอังกฤษอย่างเดียวไม่สามารถแทนเสียงในภาษาไทยได้สมบูรณ์ จะให้ต้นทุนในการแทนที่อักขระเท่ากับ C_1 ตัวอย่างเช่น สระเสียงสั้นกับสระเสียงยาว (A เท่ากับ –ะ หรือ –า)

- มาตราของอักขระไทย จะเลือกบางอักขระมาเป็นตัวแทนมาตรา ดังตาราง 4.6 โดยจะกำหนดต้นทุนในการแทนที่อักขระตัวแทนมาตรากับอักขระที่อยู่ในมาตราเดียวกันเท่ากับ C_3 ซึ่งวิธีนี้จะทำให้ความแตกต่างของเสียงได้มากกว่าวิธีจัดกลุ่มแบบซาวด์เด็กซ์ภาษาไทยที่รวมอักขระต่าง ๆ ที่อยู่มาตราเดียวกันเข้าเป็นกลุ่มเสียงเดียวกัน [13] เช่น คำว่า "สาร" กับ "ชาญ" วิธีที่นำเสนอสามารถบอกได้ว่า ส-ช และ ร-ญ ต่างกันสิ้นเชิง เนื่องจากทั้ง ส และ ช มีความสัมพันธ์กับ ด หรือ ต (มาตรา กด) แต่ ส และ ช ไม่มีความสัมพันธ์กันเอง แต่ถ้าเป็นวิธีซาวด์เด็กซ์จะบอกว่าอักษร ส และ ช มีความสัมพันธ์กันเนื่องจากอยู่ในมาตรากด และอักษร ร และ ญ มีความสัมพันธ์กันเนื่องจากอยู่ในอยู่ในมาตรากน

มาตรา	ตัวอักษร	ตัวแทนมาตรา
กก	ก ข ค ฆ	ก
กด	ค ต ถ ท ธ ฎ ฏ ฑ ฐ ฒ จ ช ศ ษ ส ซ	ด ต
กง	ง	ง
กน	น ณ ญ ล ฬ ร	น
กบ	บ ป พ ภ ฟ	บ
กม	ม	ม

ตารางที่ 4.6 อักขระแทนมาตราต่าง ๆ

- การแทนที่อักขระแบบหนึ่งอักขระต่อสองอักขระ กำหนดให้ต้นทุนในการแทนที่อักขระเท่ากับ C_1 โดยอักษรควบที่ส่งผลต่อการอ่านออกเสียงที่พิจารณาในงานวิจัยนี้คือ [14]

---

[13] นิลเนตร กรุฒวงศ์ ณ อยุธยา, "การเปลี่ยนอักขระของคำในภาษาไทย โดยใช้หลักการของซาวด์เด็กซ์" (ปริญญานิพนธ์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2534).

[14] กำชัย ทองหล่อ, หลักภาษาไทย, พิมพ์ครั้งที่ 10 (กรุงเทพมหานคร : อมรการพิมพ์, 2540).

- อักษรควบไม่แท้ มี 5 ตัว

อักษร	ออกเสียงเป็น	ตัวอย่างเช่น
ทร	ซ	ทราบ ทรง
จร	จ	จริง
สร	ส	สรง เสริม
ศร	ศ	ศรี
ซร	ซ	ไซร้

ตารางที่ 4.7 อักษรควบไม่แท้

- อักษรนำเสียงสนิท มี 9 ตัว

อักษร	ตัวอย่าง
อย	อยู่ อยาก
หง	เหงือก
หญ	ใหญ่ หญิง
หน	หนู หน่อย
หม	หมู หมอ

อักษร	ตัวอย่าง
หย	หยก หย่า
หร	หริ่ง หรูหรา
หล	หลวง หล่อ
หว	หวาน หวัง

ตารางที่ 4.8 อักษรนำเสียงสนิท

- อักษรควบที่เป็นตัวสะกด เช่น

อักษร	ตัวอย่าง
กร	จักร
ตร	จิตร
ปร	กอปร
รถ	สามารถ

อักษร	ตัวอย่าง
ดร	สมัคร
ทร	สมุทร
รค	ชลมารค
หม	พรหม

ตารางที่ 4.9 อักษรควบที่เป็นตัวสะกด

- ส่วนอื่น ๆ นอกเหนือจากนี้กำหนดต้นทุนในการแทนที่อักขระเท่ากับ C_4 (มากที่สุด)

จากแนวคิดดังกล่าวสามารถสร้างตารางการกำหนดต้นทุนในการแทนที่อักขระในส่วนของพยัญชนะและสระได้ดังแสดงเพียงบางส่วนในตารางที่ 4.10 และตารางที่ 4.11 ตามลำดับ และได้นิยาม REPLACE[a, b] หมายถึงตารางค้นหาต้นทุนในการแทนที่อักขระ b ด้วยอักขระ a โดยต้นทุนจะคำนวณจากตารางที่ 4.10 และตารางที่ 4.11

	ก	ข	ฃ	ค	...	ท	ธ	น	บ	ป	ผ	ฝ	พ	ฟ	...	พ	อ	ฮ	กร	ครู	...	หล	หว	อย
ก	C_1	C_2	C_2	C_2	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_1	C_1	...	C_4	C_4	C_4
ข	C_2	C_1	C_1	C_1	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ฃ	C_2	C_1	C_1	C_1	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ค	C_2	C_1	C_1	C_1	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
ท	C_4	C_4	C_4	C_4	...	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ธ	C_4	C_4	C_4	C_4	...	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
น	C_4	C_4	C_4	C_4	...	C_4	C_4	C_1	C_4	C_4	C_4	C_4	C_4	C_4	...	C_3	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
บ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_1	C_3	C_4	C_4	C_3	C_3	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ป	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_3	C_1	C_2	C_4	C_2	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ผ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_2	C_1	C_4	C_1	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ฝ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_1	C_4	C_1	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
พ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_3	C_2	C_1	C_4	C_1	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
ฟ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_3	C_4	C_4	C_1	C_4	C_1	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
พ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_3	C_4	C_4	C_4	C_4	C_4	C_4	...	C_1	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4
อ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_1	C_4	C_4	C_4	...	C_4	C_4	C_4
ฮ	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_1	C_4	C_4	...	C_4	C_4	C_4
กร	C_1	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_1	C_4	...	C_4	C_4	C_4
ครู	C_1	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_1	...	C_4	C_4	C_4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
หล	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_1	C_4	C_4
หว	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_1	C_4
อย	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_4	C_4	C_4	...	C_4	C_4	C_1

ตารางที่ 4.10 ตัวอย่างตารางการกำหนดต้นทุนในการแทนที่อักขระที่เป็นพยัญชนะ

<u>หมายเหตุ</u> สามารถดูรายละเอียดตารางการกำหนดต้นทุนในการแทนที่อักขระที่เป็นพยัญชนะทุกอักขระได้ในภาคผนวก ก

	ะ	ั	า	ิ	ี	ึ	ื	ุ	ู	เ	แ	โ	ใ	ไ	@	I	U	X*
ะ	C_1	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_2	C_4	C_2	C_4	C_4	C_4	C_4	C_4	C_4
ั	C_1	C_1	C_2	C_4	C_4	C_4	C_4	C_2	C_2	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4
า	C_1	C_2	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_4	C_3	C_4	C_4	C_3	C_4	C_4	C_4
ิ	C_4	C_4	C_4	C_1	C_1	C_3	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4
ี	C_4	C_4	C_4	C_1	C_1	C_4	C_4	C_4	C_4	C_2	C_2	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4
ึ	C_4	C_4	C_4	C_3	C_4	C_1	C_1	C_2	C_2	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_4
ื	C_4	C_4	C_4	C_4	C_4	C_1	C_1	C_2	C_2	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_4
ุ	C_4	C_2	C_4	C_4	C_4	C_2	C_2	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_4
ู	C_4	C_2	C_4	C_4	C_4	C_2	C_2	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_4
เ	C_4	C_4	C_4	C_4	C_2	C_4	C_4	C_4	C_4	C_1	C_2	C_4	C_4	C_4	C_4	C_4	C_3	C_4	C_4
แ	C_2	C_4	C_3	C_4	C_2	C_4	C_4	C_4	C_4	C_2	C_1	C_4	C_3	C_4	C_4	C_4	C_3	C_4	C_4
โ	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4
ใ	C_2	C_4	C_3	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_1	C_1	C_4	C_4	C_4	C_4	C_4	C_4
ไ	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_1	C_4	C_4	C_4	C_4	C_4
@	C_4	C_4	C_3	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_1	C_4	C_4	C_4	C_4
I	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_3	C_3	C_4	C_4	C_4	C_4	C_4	C_1	C_4	C_4
U	C_4	C_4	C_4	C_4	C_4	C_3	C_3	C_3	C_3	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_1	C_4
X*	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4	C_4

* X แทนรหัสสระอื่น ๆ ที่ไม่ได้แสดงในตาราง

ตารางที่ 4.11 ตารางการกำหนดต้นทุนในการแทนที่อักขระที่เป็นสระ

ส่วนการกำหนดต้นทุนในการเพิ่มหรือลบอักขระจะพิจารณาการใช้งานของอักขระในคำภาษาไทยและภาษาอังกฤษทับศัพท์ภาษาไทย โดยแบ่งเป็นกรณีต่าง ๆ ดังนี้

- สระที่มีการลดรูป ได้แก่ -ะ โ- และ -อ ซึ่งจะไม่ปรากฏในหน่วยอักษรของคำไทย แต่จะปรากฏในคำอังกฤษทับศัพท์ภาษาไทย ทำให้หน่วยอักษรที่ได้จากลดอักษรไม่ตรงกันกับคำไทย เช่น Thavee-ทวี พร-Phon และ นก-Nok เป็นต้น ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2

- พยัญชนะที่เป็นส่วนประกอบของสระประสม ได้แก่ อ ย และ ว เช่น เ-ือ เ-ีย และ -ว- เป็นต้น ซึ่งมักจะเกิดความผิดพลาดได้ง่ายในการถ่ายอักษร เช่น คำว่า Chuan (ชวน) การถอดอักษรจะได้ เชือน ซึ่งไม่ได้อักษร ว ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2
- การทับศัพท์ที่ไม่ตรงตามประกาศของราชบัณฑิตยสถาน เช่น คำว่า ธรรม-Thama จิตร-Chitara และ รัตน-Ratana เป็นต้น จากการทับศัพท์แบบนี้ทำให้เวลาถอดอักษรแล้ว จะได้สระ –ะ หรือ –า เกิน ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2

สรุปการกำหนดต้นทุนในการเพิ่มหรือลบอักขระดังตารางที่ 4.12 และได้นิยาม DELADD[a] หมายถึง ตารางค้นหาต้นทุนในการเพิ่มหรือลบอักขระ a โดยคำนวณต้นทุนจากตารางที่ 4.12

อักขระ	ต้นทุน
-ะ	C_2
เ-	C_2
อ	C_2
ย	C_2
ว	C_2
า	C_2

ตารางที่ 4.12 ตารางการกำหนดต้นทุนในการเพิ่มหรือการลบอักขระ

ส่วนฟังก์ชันกำหนดต้นทุนในการแก้ไขอักขระสามารถแบ่งออกเป็น 2 ฟังก์ชันคือ (1) การแทนที่อักขระ และ (2) การเพิ่มหรือการลบอักขระ โดยมีรายละเอียดดังนี้

การแทนที่อักขระ

การกำหนดต้นทุนในการแทนที่อักขระ โดยเปลี่ยนฟังก์ชัน $R(p_j, w_t)$ ในหัวข้อ 4.3.1 เป็น $R(p_{j-1}p_j, w_t)$ เพื่อสนับสนุนการเปรียบเทียบแบบหนึ่งอักขระต่อสองอักขระ (อักษรควบ) ซึ่งฟังก์ชันจะเลือกต้นทุนที่น้อยสุดระหว่างการเปรียบเทียบแบบหนึ่งอักขระต่อหนึ่งอักขระกับหนึ่งอักขระต่อสองอักขระเป็นคำตอบของฟังก์ชัน โดยมีรายละเอียดดังนี้

$$R(p_{j-1}p_j, w_k) = \min(\ REPLACE[p_j, w_k]\ ,\ REPLACE[p_{j-1}p_j, w_k]\ )$$

โดยที่ $p_{j-1}p_j$ เป็นอักขระสองตัวใด ๆ ที่มีตำแหน่งติดกัน

w_k เป็นอักขระใด ๆ

$REPLACE[p_j, w_k]$ ตารางค้นหาต้นทุนในการแทนที่อักขระ w_k ด้วย p_j

$REPLACE[p_{j-1}p_j, w_k]$ ตารางค้นหาต้นทุนในการแทนที่อักขระ w_k ด้วย $p_{j-1}p_j$

การเพิ่มหรือการลบอักขระ

การกำหนดต้นทุนในการเพิ่มหรือการลบอักขระ (ฟังก์ชัน $D(p_j, w_k)$ ในหัวข้อ 4.3.1) จะเลือกต้นทุนที่น้อยสุดระหว่างต้นทุนในการเพิ่มหรือลบอักขระ และต้นทุนในการแทนที่อักขระ (แทนที่สองอักขระในสายอักขระ P ด้วยหนึ่งอักขระในสายอักขระ W) เป็นคำตอบของฟังก์ชัน โดยมีรายละเอียดดังนี้

$$D(p_j, w_k) = \min(\ DELADD[p_j]\ ,\ REPLACE[p_{j-1}\ p_j, w_k]\ )$$

โดยที่ $p_{j-1}p_j$ เป็นอักขระสองตัวใด ๆ ที่มีตำแหน่งติดกัน

w_k เป็นอักขระใด ๆ

$DELADD[p_j]$ ตารางค้นหาต้นทุนในการเพิ่มหรือลบอักขระ p_j

$REPLACE[p_{j-1}\ p_j, w_k]$ ตารางค้นหาต้นทุนในการแทนที่อักขระ $p_{j-1}p_j$ ด้วย w_k

สรุปการคำนวณค่าความแตกต่างสามารถแสดงในรูปของสมการเวียนเกิด $Edit\ (P_j,\ W_k)$ แทนสมการในหัวข้อ 4.3.1 ดังนี้

$$Edit(P_0, W_0) = 0$$

$$Edit(P_j, W_0) = Edit(P_{j-1}, W_0) + D(p_{j-1}, p_j)$$

$$Edit(P_0, W_k) = Edit(P_0, W_{k-1}) + D(w_{k-1}, w_k)$$

$$Edit(P_j, W_k) = \min[\ Edit(P_{j-1}, W_k) + D(p_{j-1}, w_j),$$
$$Edit(P_j, W_{k-1}) + D(p_{k-1}, w_k),$$
$$Edit(P_{j-1}, W_{k-1}) + R(p_{j-1}p_j, w_k)\ ]$$

โดยที่

$P_j = p_1p_2p_3...p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1w_2w_3...w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$p_{j-1}p_j$ เป็นสองอักขระใด ๆ ที่มีตำแหน่งติดกัน

$D(p_j, w_k)$ มีค่าเท่ากับ min(DELADD[p_j] , REPLACE[p_{j-1} p_j , w_k])

$R(p_{j-1}p_j, w_k)$ มีค่าเท่ากับ min(REPLACE[p_j, w_k] , REPLACE[$p_{j-1}p_j$, w_k])

4.3.2 เงื่อนไขในการเปรียบเทียบรหัสคำ

เงื่อนไขที่ใช้ในการทดสอบว่าคู่รหัสคำที่นำมาเปรียบเทียบเป็นรหัสคำที่ได้มากจากคำหลักที่ตรงกันระหว่างภาษาไทยกับภาษาอังกฤษทับศัพท์ภาษาไทยหรือไม่นั้น มีดังนี้

$$Edit(P^c_m, W^c_n) \leq \alpha \times Max(Len(P^c_m) , Len(W^c_n)) \times C_4 \text{ และ}$$

$$Edit(P^v_m, W^v_n) \leq \alpha \times Max(Len(P^v_m) , Len(W^v_n)) \times C_4$$

โดยที่

- P^c_m คือรหัสคำ P เฉพาะในส่วนของพยัญชนะ มีความยาว m ตัวอักษร
- P^v_m คือรหัสคำ P เฉพาะในส่วนของสระ มีความยาว m ตัวอักษร
- Edit(P, W) ค่าความแตกต่างของรหัสคำ P กับ รหัสคำ W
- α (แอลฟา) คือพารามิเตอร์ที่ปรับประสิทธิผล มีค่าระหว่าง 0 – 1
- Max(j, k) คือฟังก์ชันที่จะเลือกค่าที่มากที่สุดระหว่าง j กับ k
- Len(P) ความยาวของรหัสคำ P
- C_4 คือต้นทุนในการแก้ไขอักขระที่มากที่สุด

จากเงื่อนไขดังกล่าวในส่วนของ Max(Len(P), Len(W)) × C_4 หมายถึงต้นทุนมากสุดที่ใช้ในการแก้ไขอักขระคือแก้ไขทุกอักขระของ W ให้เท่ากับ P ส่วนค่าของแอลฟานี้เป็นพารามิเตอร์ของระบบที่ผู้ใช้สามารถกำหนดได้ ซึ่งจะส่งผลต่อค่าแม่นยำและค่าเรียกคืนของระบบ ค่าแอลฟามีค่าระหว่าง 0 ถึง 1 ซึ่งจะเป็นตัวกำหนดเกณฑ์การยอมรับความเหมือนกันของรหัสคำแบบ

ประมาณ โดยที่ 0 หมายถึงรหัสคำทั้งสองต้องเหมือนกันทุกประการจึงจะยอมรับ ในขณะที่ 1 หมายถึงการยอมรับทุก ๆ คู่รหัสคำไม่ว่าจะแตกต่างเท่าไรก็ตาม

รหัสคำที่ได้มาจากคำหลักภาษาไทยกับภาษาอังกฤษทับศัพท์ภาษาไทยนั้นจะตรงกันก็ต่อเมื่อ สมการเงื่อนไขต้องเป็นจริงทั้งส่วนของพยัญชนะ และส่วนของสระ

ตัวอย่าง การทดสอบคำว่า "CHULERTTIYAWONG" และ "ชูเลิศติยวงศ์" เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่ โดยจะเริ่มต้นจากการเข้ารหัสคำ การคำนวณหาค่าความแตกต่าง และนำค่าความแตกต่างที่ได้ไปทดสอบกับเงื่อนไขเพื่อสรุปว่าเป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่

<u>การเข้ารหัสคำ</u>

- CHULERTTIYAWONG → ชุกqดติยะวอง → ชกดดยวอง
- ชูเลิศติยวงศ์ → ชูเลิศติยวง → ชูกqศติยวง → ชกศดยวง

<u>การคำนวณหาค่าความแตกต่าง</u>

- ส่วนพยัญชนะของรหัสคำ Edit(ชกดดยวอง, ชกศดยวง) = 5
- ส่วนสระของรหัสคำ Edit(,) = 1

<u>การทดสอบเงื่อนไขในการเปรียบเทียบรหัสคำ</u>

กำหนดให้ $\alpha = 0.15$

$C_4 = 7$

$Edit(P_m^c, W_n^c) \leq \alpha \times Max(Len(P_m^c), Len(W_n^c)) \times C_4$ และ

$Edit(P_m^v, W_n^v) \leq \alpha \times Max(Len(P_m^v), Len(W_n^v)) \times C_4$

Edit(ชกดดยวอง, ชกศดยวง) $\leq 0.15 \times$ Max(Len(ชกดดยวอง), Len(ชกศดยวง)) $\times 7$ และ

Edit(,) $\leq 0.15 \times$ Max(Len( ), Len( )) $\times 7$

$4 \leq 0.15 (Max(8, 7) \times 7)$ และ

$1 \leq 0.15 (Max(4, 3) \times 7)$

$4 \leq 0.15 (8 \times 7)$ และ

$1 \leq 0.15 (4 \times 7)$

$4 \leq 8.4$ และ

$1 \leq 4.2$

จากตัวอย่างพบว่า $4 \leq 8.4$ และ $1 \leq 4.2$ เป็นจริง เพราะฉะนั้นขั้นตอนวิธีสรุปคำศัพท์ทั้งสองคำเป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษ

4.4 วิธีการทดลอง

ผู้วิจัยได้ทำการทดลองขั้นตอนวิธีที่ได้นำเสนอโดยใช้ชื่อและชื่อสกุลทั้งภาษาไทยและภาษาอังกฤษที่ตรงกันของนิสิตปัจจุบัน ระดับปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย จำนวน 5,000 คู่ เป็นตัวอย่างข้อมูลในการทดลอง โดยนำคำศัพท์ทั้งหมดไปทำการเข้ารหัสด้วยขั้นตอนวิธีที่นำเสนอในหัวข้อ 4.2 และจัดเก็บคำศัพท์และรหัสคำในฐานข้อมูล หลังจากนั้นนำคำศัพท์ทั้งหมด 10,000 คำ ไปค้นคืนทีละคำศัพท์กับฐานข้อมูลด้วยขั้นตอนวิธีการเปรียบเทียบรหัสคำที่ได้นำเสนอในหัวข้อ 4.3 เพื่อทำการคำนวณค่าแม่นยำ และค่าเรียกคืนสำหรับการค้นคืนใช้สูตรดังนี้

$$\text{ค่าแม่นยำ} = \frac{\text{จำนวนคำศัพท์ที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำศัพท์ที่คืนกลับมา}} \times 100$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำศัพท์ที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำศัพท์ที่เกี่ยวข้องทั้งหมด}} \times 100$$

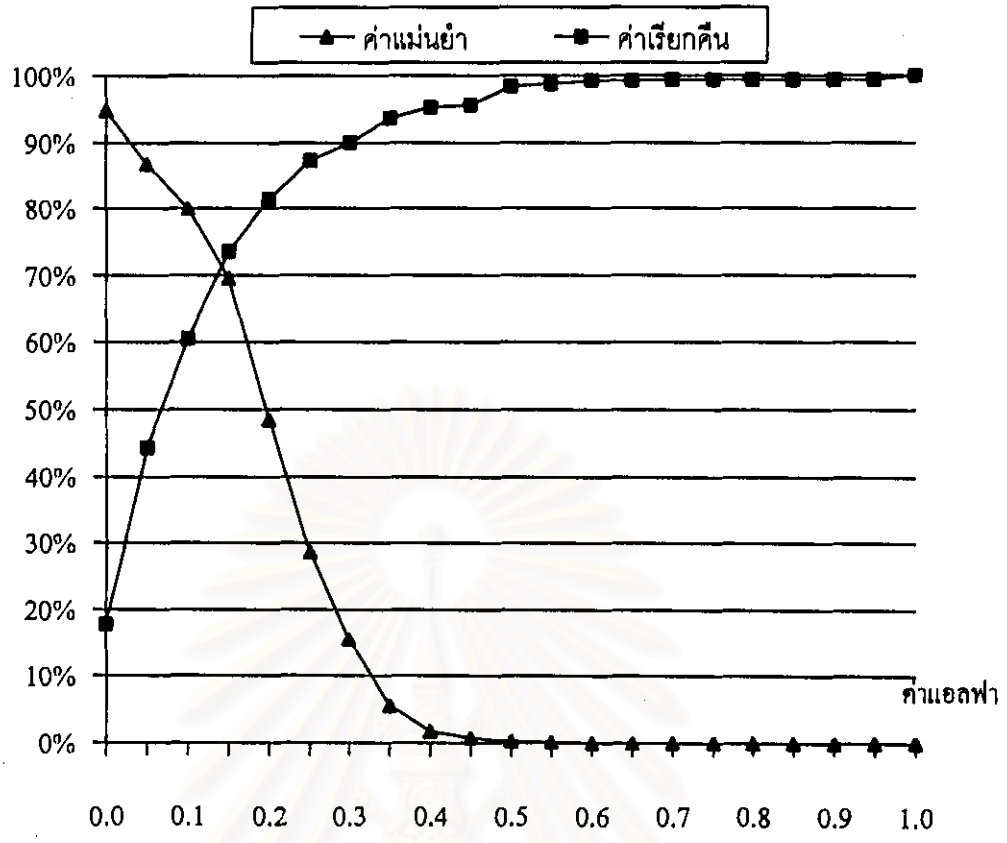
ผู้วิจัยได้ทำการทดลองโดยการใช้ค่าแอลฟาในระดับต่าง ๆ เพื่อหาความสัมพันธ์ระหว่างค่าแอลฟากับประสิทธิผลของระบบค้นคืน

4.5 ผลการทดลอง

จากการทดลองในหัวข้อ 4.4 เพื่อหาความสัมพันธ์ระหว่างค่าแอลฟากับประสิทธิผลของระบบค้นคืนโดยการเปลี่ยนแปลงค่าแอลฟาทุก ๆ 0.05 จาก 0.00 ถึง 1.00 ได้ผลแสดงในตารางที่ 4.13 และรูปที่ 4.4

ค่าแอลฟา	ค่าแม่นยำ	ค่าเรียกคืน
0.00	0.947752	0.177200
0.05	0.867973	0.442200
0.10	0.800750	0.605800
0.15	0.694093	0.734800
0.20	0.484050	0.814800
0.25	0.286261	0.872800
0.30	0.154865	0.899000
0.35	0.056202	0.936200
0.40	0.018329	0.952800
0.45	0.007337	0.955800
0.50	0.002760	0.984000
0.55	0.001368	0.986800
0.60	0.000750	0.992200
0.65	0.000517	0.992200
0.70	0.000348	0.993600
0.75	0.000271	0.993600
0.80	0.000235	0.993600
0.85	0.000222	0.993600
0.90	0.000214	0.993600
0.95	0.000211	0.993600
1.00	0.000200	1.000000

ตารางที่ 4.13 ความสัมพันธ์ระหว่างค่าแอลฟากับประสิทธิผลของระบบค้นคืน



รูปที่ 4.4 ความสัมพันธ์ระหว่างค่าแอลฟากับประสิทธิผลของระบบค้นคืน

จากรูปที่ 4.4 แสดงให้เห็นว่าค่าแม่นยำของระบบค้นคืนสูงประมาณ 94 เปอร์เซ็นต์ และจะลดค่าลงอย่างต่อเนื่อง เมื่อค่าของแอลฟาเพิ่มขึ้น ส่วนพฤติกรรมของค่าเรียกคืนจะเริ่มต้นประมาณ 17 เปอร์เซ็นต์และจะเพิ่มค่าขึ้นเมื่อค่าแอลฟาเพิ่มขึ้น ประสิทธิผลของระบบที่ได้จากการทดลองพบว่า สอดคล้องกับพฤติกรรมโดยปกติของค่าแม่นยำและค่าเรียกคืนในระบบค้นคืนใด ๆ คือแนวโน้มของค่าแม่นยำและค่าเรียกคืนจะตรงข้ามกัน

จากผลการทดลองพบว่าขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษจะเกิดประสิทธิผลสูงสุดคือค่าแม่นยำเท่ากับ 69 เปอร์เซ็นต์และค่าเรียกคืนเท่ากับ 73 เปอร์เซ็นต์ เมื่อกำหนดค่าแอลฟาเท่ากับ 0.15

4.6 สรุป

ในบทนี้ได้กล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษแบบภาษาอังกฤษทับศัพท์ภาษาไทย การค้นคืนข้ามภาษานั้นสามารถทำได้โดยนำข้อคำถามไปเข้ารหัสคำแล้วนำรหัสคำที่ได้ไปค้นหาในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี

ขั้นตอนวิธีที่ได้นำเสนอนี้ ผู้วิจัยได้ใช้หลักภาษาไทยมาช่วยในการออกแบบการประมวลผลตัวอักษรเบื้องต้นเพื่อให้คำศัพท์ทั้งสองภาษามีรูปแบบที่เหมือนกันก่อนที่ไปนำไปเข้ารหัสคำ หลังจากนั้นได้ใช้เทคนิคระยะแก้ไขการอ่านออกเสียงสั้นที่สุดมาดัดแปลงเพื่อให้เหมาะสมกับการอ่านออกเสียงของภาษาไทยเพื่อใช้ในการเปรียบเทียบกับคู่รหัสคำที่ได้ และได้โดยการออกแบบเงื่อนไขที่ใช้ทดสอบว่ารหัสคำทั้งสองว่าเป็นรหัสคำที่ได้มาจากคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษ

ผลการทดลองขั้นตอนวิธีที่นำเสนอ พบว่าค่าแม่นยำสูงถึง 69 เปอร์เซ็นต์ และค่าเรียกคืนสูงถึง 73 เปอร์เซ็นต์ เมื่อกำหนดให้ค่าแอลฟ่าเท่ากับ 0.15