**บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง**

เนื้อหาในส่วนนี้นำเสนอทฤษฎีและงานวิจัยที่เกี่ยวข้องกับ การปริวรรตอักษรด้วยเครื่องสำหรับภาษาสันสกฤตแบบโรมันเป็นไทย มีดังต่อไปนี้

**2.1 ทฤษฎีที่เกี่ยวข้อง**

**2.1.1 การปริวรรตอักษรหรือการเขียนคำทับศัพท์ (transliteration)**

การปริวรรตอักษร คือการดำเนินการแปลงอักษรหรืออักขรวิธีจากระบบการเขียนหรือภาษาหนึ่งไปสู่อีกระบบหนึ่งอย่างมีหลักการ เมื่อปริวรรตแล้วสามารถแปลงกลับเป็นอักษรหรืออักขรวิธีเดิมได้[16] เพื่อให้สามารถเขียนคำในภาษาต่างประเทศด้วยภาษาและอักษรในภาษานั้น ๆ ได้สะดวก เช่น การปริวรรตอักษรโรมันภาษาอังกฤษ มาเป็นอักษรไทยเพื่อใช้ในภาษาไทย หรือการปริวรรตอักษรภาษาไทย ไปเป็นอักษรโรมันเพื่อใช้ในภาษาอังกฤษ เป็น ปกติแล้วการปริวรรตอักษรคือการจับคู่จากระบบการเขียนหนึ่งไปยังอีกระบบหนึ่งแบบคำต่อคำหรืออักษรต่ออักษร การปริวรรตอักษรได้พยายามที่จะสร้างความสัมพันธ์หนึ่งต่อหนึ่งและทำให้เกิดความถูกต้องแม่นยำ เพื่อให้ผู้อ่านที่ได้รับรู้สามารถสะกดคำต้นฉบับจากคำปริวรรตอักษรได้ ดังนั้นจึงมีการกำหนดหลักการปริวรรตอักษรที่ซับซ้อนในการจัดการกับตัวอักษรบางตัวในภาษาต้นฉบับที่ไม่สัมพันธ์กับอักษรในภาษาเป้าหมาย ความหมายอย่างแคบของการปริวรรตอักษรคือ การปริวรรตอักษรแบบถอดอักษร (transliteration) ซึ่งเคร่งครัดในการคงตัวอักษรและเครื่องหมายวรรคตอนทุกอย่างเอาไว้ ทั้งนี้การถอดอักษรไม่สนใจความแตกต่างของเสียงในภาษา ตัวอย่างหนึ่งของการถอดอักษรคือการใช้แป้นพิมพ์ภาษาอังกฤษ พิมพ์แทนภาษาอื่นที่ใช้ตัวอักษรต่างออกไปเช่นภาษารัสเซียเนื่องจากมีข้อจำกัดทางเทคนิค หรือการถอดอักษรโบราณเพื่อให้ยังคงรักษารูปแบบการเขียนเดิมเอาไว้การปริวรรตอักษรเป็นการถอดอักษร ต่างจากการถอดเสียง (transcription) ซึ่งเป็นการจับคู่เสียงอ่านของภาษาหนึ่ง ๆ ไปยังรูปแบบการเขียนของอีกภาษาที่ใกล้เคียงที่สุด ถึงแม้ว่าระบบการถอดอักษรส่วนใหญ่จะยังคงจับคู่อักษรต้นฉบับกับอักษรในภาษาเป้าหมายที่ออกเสียงคล้ายกันในบางคู่ ถ้าหากความสัมพันธ์ระหว่างตัวอักษรกับเสียงเหมือนกันทั้งสองภาษา การถอดอักษรก็อาจแทบจะเหมือนกับการถอดเสียง

**2.1.2 ภาษาสันสกฤต**

ภาษาสันสกฤตเป็นภาษาที่เก่าแก่ที่สุดภาษาหนึ่งในภาษากลุ่มอินโด-ยูโรเปียน (Indo-European) สาขาย่อยอินโด-อิเรเนียน (Indo-Iranian) และอยู่ในกลุ่มย่อยอินโด-อารยัน (Indo-Aryan) โดยมีระดับวิวัฒนาการเก่าแก่ในระดับใกล้เคียงกับภาษาในกลุ่มอินโด-ยูโรเปียนด้วยกัน คือภาษาละตินและภาษากรีกโบราณ วรรณกรรมภาษาสันสกฤตพบการใช้ที่หลากหลาย ตั้งแต่เป็นวรรณคดี บทกวี บทละคร เป็นตำราทางวิชาการหลากหลายสาขา และเป็นใช้ทางศาสนา บันทึกบทสวด ปรัชญา หลักการทางศาสนา ทั้งในพบเอกสารทั้งใน ศาสนาพราหมณ์-ฮินดู พุทธ เชน และ ซิกข์ ซึ่งในส่วนของ ศาสนาพราหมณ์ – ฮินดูถือว่า ภาษาสันสกฤตเป็นภาษาศักดิ์สิทธิ์

คำว่า สันสกฤต แปลว่า “กลั่นกรองแล้ว” ซึ่งเป็นภาษาของชนชั้นพราหมณ์ ตรงข้ามกับภาษาพูดของชาวบ้านทั่วไปที่เรียกว่า “ปรากฤต” ภาษาสันสกฤตมีพัฒนาการในหลายยุคสมัย โดยมีหลักฐานเก่าแก่ที่สุด ภาษา ภาษาพระเวท (Vedic Sanskrit) ที่ปรากฏในคัมภีร์ฤคเวท เมื่อราว 1,200 ปีก่อน ค.ศ. อันเป็นบทสวดสรรเสริญพระเจ้าในลัทธิพราหมณ์ในยุคต้น ภาษาพระเวทดั้งเดิมยังมิได้มีการวางกฎเกณฑ์ให้เป็นระเบียบรัดกุมและสละสลวย และมีหลักทางไวยากรณ์อย่างกว้าง ๆ ราว 57 ปีก่อน พ.ศ. พราหมณ์ชื่อ "ปาณินิ" ชาวแคว้นคันธาระ ท่านเห็นว่าภาษาสันสกฤตแบบพระเวทนั้นมีภาษาถิ่นปนเข้ามา หากไม่เขียนไวยากรณ์ที่เป็นระเบียบแบบแผนไว้จะคละกับภาษาถิ่น ปาณินิได้ศึกษาและจัดเรียบเรียงตำราไวยากรณ์ขึ้น 8 บท ชื่อว่า “อัษฏาธยายี” ภาษาที่ปรับปรุงใหม่นี้เรียกว่า "ตันติสันสกฤต" หรือ สันสกฤตแบบแผน (Classical Sanskrit) วรรณคดีสันสกฤตแบบแผนที่สำคัญและเป็นที่รู้จักมากคือ มหาภารตะ และ รามายณะ ภาษาสันสกฤตอีกสาขาหนึ่ง เรียกว่า ภาษาสันสกฤตผสม หรือ ภาษาสันสกฤตผสมในพุทธศาสนา (Buddhist Hybrid Sanskrit or Mixed Sanskrit) เป็นภาษาสันสกฤตยุคหลังถัดจากภาษาสันสกฤตแบบแผน พบในคัมภีร์ทางพระพุทธศาสนาทั้งในนิกายสรรวาสติวาทและพระพุทธศาสนาฝ่ายมหายาน วรรณกรรมภาษาสันสกฤตพบการใช้ที่หลากหลาย มีพัฒนาการมาอย่างต่อเนื่องสามารถ แบ่งประเภทตามเนื้อหาได้เป็น 4 ประเภทใหญ่ๆคือ

1. อาคม(āgama) มีเนื้อหาเกี่ยวกับศาสนาและปรัชญา
2. อิติหาส(itihāsa) มีเนื้อหาเกี่ยวกับประวัติศาสตร์หรือวีรชนและประเพณีที่สืบทอดกันมา
3. ศาสตร์(śāstra) มีเนื้อหาเกี่ยวกับศิลปวิทยาการและงานวิชาการ
4. กาวยะ(kāvya) กวีนิพนธ์หรือบทประพันธ์ที่อยู่ในรูปของศิลปะ

ภาษาสันสกฤตไม่มีอักษรสำหรับเขียนชนิดใดชนิดหนึ่งโดยเฉพาะ เป็นภาษาที่ไม่มีระบบการเขียนเป็นของตนเอง แต่จะเขียนด้วยอักษรหลายชนิด อักษรเก่าแก่ที่ใช้เขียนภาษาสันสกฤตมีหลายชนิดด้วยกัน เช่น อักษรขโรษฐี อักษรพราหมี อย่างไรก็ตามในปัจจุบันโดยทั่วไปนิยมเขียนภาษาสันสกฤตด้วยอักษรเทวนาครี ส่วนอักษรอื่น ๆ แล้วแต่ความนิยมในแต่ละท้องถิ่นในอินเดีย ทั้งนี้เนื่องจากอักษรที่ใช้ในอินเดีย มักจะเป็นตระกูลเดียวกัน จึงสามารถดัดแปลงและถ่ายทอด (Transliteration) ระหว่างชุดอักษรได้ง่าย แม้กระทั่งในเอเชียตะวันออกเฉียงใต้มีชุดอักษรท้องถิ่นที่ใช้เขียนภาษาสันสกฤตได้ ยังมีจารึกโบราณภาษาสันสกฤตที่ใช้ อักษรปัลลวะ อักษรขอมโบราณ ในระดับสากลยังใช้อักษรโรมันเขียนภาษาสันสกฤตโดยมีมาตรฐานแตกต่างกันออกไปด้วย ตัวอย่างลักษณะการบันทึกแบบอักษรต่างๆ ดังที่แสดงในรูปที่ 2.1,2.2,2.3



รูปที่ 2.1 ตัวอย่างการเขียนภาษาสันสกฤตด้วยอักษรชนิดต่างๆ

(ที่มา: wikipedia.org, ปี พ.ศ. 2558)



รูปที่ 2.2 ตัวอย่างการเขียนภาษาสันสกฤตด้วยอักษรโรมันแบบต่างๆ



รูปที่ 2.3 ตัวอย่างจารึกภาษาสันสกฤตในประเทศไทย   
 จารึกวัดมเหยงค์ เป็นจารึกอักษรปัลลวะ อายุพุทธศตวรรษที่ ๑๒

(ที่มา: ศูนย์มานุษยวิทยาสิรินธร (องค์การมหาชน), ปี พ.ศ. 2551)

**2.1.3 การปริวรรตอักษรภาษาสันสกฤตสากล (IAST)**

การปริวรรตอักษรภาษาสันสกฤตสากล หรือ The International Alphabet of Sanskrit Transliteration (IAST) เป็นรูปแบบการปริวรรตอักษรภาษาสันสกฤตเป็นอักษรโรมัน(Romanization) รูปแบบหนึ่งที่ปราศจากการสูญเสีย(Lossless) คือสามารถจับคู่อักษรต้นทางและปลายทางได้ครบคู่ โดยมากอักษรต้นทางนั้นมักเป็นอักษรอินเดียตระกูลต่างๆ(Indic Script) นอกจากใช้ปริวรรตอักษรภาษาสันสกฤตให้เป็นอักษรโรมัน ยังนิยมใช้ปริวรรตอักษรภาษาปรากฤต อื่นๆเช่น ภาษาบาลี และภาษาอปภรัมศะ เป็นต้น

แต่เดิมในการการปริวรรตอักษรภาษาสันสกฤตเป็นอักษรโรมันยังไม่ได้มีมาตรฐานกลางที่ใช้ร่วมกันแต่ใช้วิธีการปริวรรตตามข้อเสนอของผู้เชี่ยวชาญหลายๆท่าน ดังที่แสดงในรูปที่ 2.4



รูปที่ 2.4 การปริวรรตอักษรภาษาสันสกฤตเป็นอักษรโรมันตามข้อเสนอของบุคคล   
 ต่างๆก่อนมีแบบการปริวรรตอักษรภาษาสันสกฤตสากล[13]

(ที่มา: shashir.autodidactus.org, ปี พ.ศ. 2555)

จนกระทั่งในปี ค.ศ. 1894 มีการประชุมทางวิชาการเกี่ยวกับด้านตะวันออกศึกษาหรือเอเชียศึกษาในปัจจุบัน การประชุมผู้เชี่ยวชาญด้านตะวันออกศึกษานานาชาติ ครั้งที่ 10 ที่นครเจนีวา ประเทศสวิสแลนด์ (10th International Congress of Orientalists, Held at Geneva) มีมติที่ประชุมให้รวมรูปแบบการปริวรรตอักษรภาษาสันสกฤตและบาลีเป็นอักษรโรมัน จากสองรูปแบบหลัก 2 รูปแบบคือ รูปแบบการปริวรรตของ ราชสมาคมเอเชียแห่งบริเตนใหญ่และไอร์แลนด์ (Royal Asiatic Society of Great Britain and Ireland) [14]และรูปแบบของสมาคมตะวันออกศึกษาแห่งเยอรมัน (German Oriental Society : Deutsche Morgenländische Gesellschaft) และตีพิมพ์สรุปรายงานการประชุมในปีเดียวกันเป็นภาษาฝรั่งเศสลงในหนังสือ Xme Congrès International des Orientalistes, Session de Genève. Rapport de la Commission de Transcription (1894) จากนั้นในปี ค.ศ.1895 ราชสมาคมเอเชียแห่งบริเตนใหญ่และไอร์แลนด์ได้แปลสรุปรายงานการประชุมเป็นภาษาอังกฤษลงในวารสารราชสมาคมเอเชีย [15] ต่อมารูปแบบการปริวรรตนี้ มีความสำคัญทางวิชาการภาษาสันสกฤตมาก  
 จึงได้เป็น “การปริวรรตอักษรภาษาสันสกฤตสากล” หรือ “The International Alphabet of Sanskrit Transliteration” (IAST) เพราะเป็นมาตรฐานหลักการปริวรรตอักษรภาษาสันสกฤตเป็นอักษรโรมัน (Romanization) จนถึงปัจจุบัน



รูปที่ 2.4 บางส่วนในรายงานมติที่ประชุม เรื่องการปริวรรตอักษรภาษาสันสกฤตและบาลี   
 สากล ในการประชุมผู้เชี่ยวชาญด้านตะวันออกศึกษานานาชาติ ครั้งที่ 10 [15]

(ที่มา: วารสารราชสมาคมเอเชีย, ปี ค.ศ.1985)

**2.1.4 การปริวรรตอักษรภาษาสันสกฤตแบบโรมัน IAST เป็นไทย**

การปริวรรตอักษรภาษาสันสกฤตแบบโรมัน IAST เป็นไทยนั้นผู้ปริวรรตต้องมีความรู้ความเข้าใจเกี่ยวกับรูปแบบอักษรในแต่ละส่วนมีตำแหน่งละหน้าที่อย่างไรจึงจะปริวรรตอักษรได้อย่างถูกต้อง สำหรับการเปรียบเทียบรูปอักษรโรมัน IAST เป็นไทย มีส่วน ประกอบดังต่อไป 4 ส่วนดังนี้

1. รูปสระ(Vowels) แบ่งได้ 2 ประเภท คือ
   1. สระแท้ แสดงในตารางที่ 2.1
   2. สระประสม แสดงในตารางที่ 2.1
2. รูปพยัญชนะ(Consonants) แบ่งได้ 2 ประเภท คือ
   1. พยัญชนะวรรค แสดงในตารางที่ 2.2
   2. พยัญชนะอวรรค แสดงในตารางที่ 2.3
3. รูปเครื่องหมายและสัญลักษณ์พิเศษ(Diacritics) แสดงในตารางที่ 2.4
4. รูปตัวเลข(Numerals) รูปอักษรโรมัน IAST ใช้ตัวเลขฮินดู-อารบิก แสดงในตารางที่ 2.5

ตารางที่ 2.1 การเปรียบเทียบรูปสระ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ประเภท** | **ฐานที่เกิดเสียง** | **หรัสวะ**  **(สระเสียงสั้น)** | **ทีรฆะ (สระเสียงยาว)** | |
| **สระแท้**  **สมานากษะระ** | กัณฐยะ (คอ) | A a อะ | Ā ā อา | |
| ตาละวยะ(เพดาน) | I i อิ | Ī ī อี | |
| มูรธันยะ(ปุ่มเหงือก) | U u อุ | Ū ū อู | |
| ทันตยะ(ฟัน) | Ṛ ṛ ฤ | Ṝ ṝ ฤๅ | |
| โอษฐยะ(ริมผีปาก) | Ḷ ḷ ฦ | Ḹ ḹ ฦๅ | |
| **สระประสม**  **(สันธยักษะระ)** | กัณฐะตาละวยะ(คอและเพดาน) |  | E e เอ | AI ai ไอ |
| กัณโฐษฐยะ(คอและริมฝีปาก) |  | O o โอ | AU au เอา |

ตารางที่ 2.2 การเปรียบเทียบรูปพยัญชนะวรรค

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **วรรค** | **ฐานที่เกิดเสียง** | **สปรรศะ(เสียงกัก)** | | | | **อนุนาสิกะ (เสียงขึ้นจมูก)** |
| **อโฆษะ(เสียงไม่ก้อง)** | | | **โฆษะ(เสียงก้อง)** | |
| **ลมเบา\*** | **ลมหนัก\*\*** | **ลมเบา\*** | **ลมหนัก\*\*** | |
| ก วรรค | กัณฐยะ (คอ) | K k  ก | KH kh  ข | G g  ค | gh  ฆ | Ṅ ṅ  ง |
| จ วรรค | ตาละวยะ(เพดาน) | C c  จ | CH ch  ฉ | J j  ช | JH jh  ฌ | Ñ ñ  ญ |
| ฎ วรรค | มูรธันยะ(ปุ่มเหงือก) | Ṭ ṭ  ฏ | ṬH ṭh  ฐ | Ḍ ḍ  ฑ | ḌH ḍh  ฒ | Ṇ ṇ  ณ |
| ต วรรค | ทันตยะ(ฟัน) | T t  ต | TH th  ถ | D d  ท | DH dh  ธ | N n  น |
| ป วรรค | โอษฐยะ(ริมผีปาก) | P p  ป | PH ph  ผ | B b  พ | BH bh  ภ | M m  ม |

*\* ลมเบา คือ อัลปปราณะ \*\* ลมหนัก คือ มหาปราณะ*

ตารางที่ 2.3 การเปรียบเทียบรูปพยัญชนะอวรรค

|  |  |  |  |
| --- | --- | --- | --- |
| **ฐานที่เกิดเสียง** | **อันตะสถะ(เสียงเปิด)** | **อูษมะ(เสียงเสียดแทรก)** | |
| **โฆษะ(ก้อง)** | **อโฆษะ(ไม่ก้อง)** | **โฆษะ(ก้อง)** |
| **อัลปปราณะ(ลมเบา)** | **มหาปราณะ(ลมหนัก)** | |
| กัณฐยะ(คอ) |  |  | H h  ห |
| ตาละวยะ(เพดาน) | Y y  ย | Ś ś  ศ |  |
| มูรธันยะ(ปุ่มเหงือก) | R r  ร | Ṣ ṣ  ษ |  |
| ทันตยะ(ฟัน) | L l  ล | S s  ส |  |
| โอษฐยะ(ริมผีปาก) | V v  ว |  |  |
| พยัญชนะ อวรรคอื่นๆ | Ḻ ḻ  ฬ \* | | |

*\* มีในภาษาสันสกฤตแบบพระเวท และภาษาบาลี*

ตารางที่ 2.4 การเปรียบเทียบรูปเครื่องหมายและสัญลักษณ์พิเศษ

|  |  |  |  |
| --- | --- | --- | --- |
| **ชื่อสัญลักษณ์** | **คำอธิบาย** | **สัญลักษณ์** | **ปริวรรตเป็นอักษรไทย** |
| อนุสวาระ | เครื่องหมายกำกับสระใช้กำกับเสียงขึ้นจมูก | Ṃ ṃ | ˚ (นิคหิต) |
| วิสรรคะ | เครื่องหมายกำกับสระ ใช้เพื่อแทนเสียงลมหายใจมาก | Ḥ ḥ | ะ (วิสรรชนีย์) |
| อวครหะ | เครื่องหมายกำกับ ว่า สระอะมีการย่อหรือตัดออกในการสนธิในภาษาสันสกฤต | ' (apostrophe) | ’ (อัญประกาศเดียว) |
| จันทรพินทุ/ อนุนาสิกะ | เครื่องหมายกำกับ บนเสียงอนุสวาระ ที่เน้นเสียงพิเศษคือ เน้นเสียงขึ้นจมูกกว่าปกติ | M̐ m̐ ,  Ṁ ṁ\* | ัํ มีผู้ใช้ไม้หันอากาศและมีนิคหิตข้างบน เช่น dasyūm̐radharām̐ ทสฺยูัํรธรัํา \*\* |
| ทัณฑะ | เครื่องหมายจบประโยคหรือจบบาทที่สองของโศลก | । | ฯ อังคั่นเดี่ยว |
| ทวิทัณฑะ | จบประโยคเพื่อจะขึ้น  ย่อหน้าใหม่หรือจบโศลก | ॥ | ๚ อังคั่นคู่ |

*\* ประกาศสำนักนายกรัฐมนตรี เรื่อง หลักเกณฑ์การทับศัพท์ภาษาจีนและภาษาฮินดี ลงวันที่ 22 ธันวาคม 2549 ใช้จันทรพินทุเป็น Ṁ ṁ* [17]

*\*\*ปริวรรตเป็นอักษรไทยสำหรับ จันทรพินทุ มีการกำหนดเฉพาะ ภาษาฮินดี ตามประกาศสำนักนายกรัฐมนตรีข้างต้นให้ใช้ “ง”* [17] *ส่วนการปริวรรตจากภาษาสันสกฤตนั้นยังไม่มีการกำหนดจากหน่วยงานที่เกี่ยวข้องอย่างเป็นทางการ*

ตารางที่ 2.5 การเปรียบเทียบรูปตัวเลข

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ตัวเลขฮินดู-อารบิก** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **ตัวเลขไทย** | ๐ | ๑ | ๒ | ๓ | ๔ | ๕ | ๖ | ๗ | ๘ | ๙ |

**2.1.4 ทฤษฎีการปริวรรตอักษรด้วยเครื่อง**

การแปลภาษาด้วยเครื่องสามารถแบ่งตามลักษณะการทำงานของระบบออกเป็น 3 กลุ่ม คือ

2.2.1 การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ (Rule-based Machine Translation) เป็นการแปลภาษาด้วยเครื่องโดยใช้ความรู้ทางด้านภาษาศาสตร์มาใช้ในการกำหนดกฎเกณฑ์ของระบบ ซึ่งจะต้องมีการแยกคุณลักษณะและข้อมูลทางภาษาศาสตร์ของภาษาต้นฉบับจากนั้นจึงทำการวิเคราะห์ตามกฎไวยากรณ์ของภาษาต้นฉบับ แล้วส่งผ่านข้อมูลที่ได้จากการวิเคราะห์นั้นไปยังกระบวนการวิเคราะห์คุณลักษณะและข้อมูลทางภาษาศาสตร์ของภาษาเป้าหมายแล้วจึงทำการแปลจากพจนานุกรมคู่ภาษาและสร้างรูปประโยคของภาษาเป้าหมายออกมา ข้อจำกัดของการแปลภาษาด้วยเครื่องกลุ่มนี้คือ ผู้พัฒนาต้องมีความรู้ทางด้านภาษาศาสตร์ ของทั้งสองภาษาเป็นอย่างมาก จึงจะทำให้ได้การแปลที่มีประสิทธิภาพ

2.2.2 การแปลภาษาด้วยเครื่องแบบใช้สถิติ (Statistical Machine Translation)เป็นการแปลภาษาด้วยเครื่องโดยใช้วิธีการทางสถิติมาช่วยในการแปล ซึ่งการแปลกลุ่มนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีการจับคู่ประโยค เพื่อเป็นฐานความรู้ให้ระบบทำการเรียนรู้ และใช้ค่าทางสถิติ เอ็นแกรม (N-Gram Model) ซึ่งเป็นการคำนวณค่าของการที่คำเกิดขึ้นร่วมกัน ถ้าชุดคำชุดใดมีค่าเอ็นแกรมหรือค่าความน่าจะเป็น (Probability) สูง แสดงว่าชุดคำนี้มีโอกาสเกิดขึ้นร่วมกันบ่อยครั้ง การคำนวณหาค่าเอ็นแกรมของชุดคำที่มีอยู่ในฐานข้อมูลคู่ภาษา ทำให้ได้ค่าความน่าจะเป็นของชุดคำต่างๆ เพื่อนำไปเปรียบเทียบและใช้ในการแปลข้อความได้ โดยสามารถเลือกใช้ค่าเอ็นแกรมได้ตั้งแต่ 2 คำ (bigrams), 3 คำ (trigrams) เป็นต้น ยิ่งใช้ค่าเอ็นแกรมมาก ยิ่งทำให้ภาษาต้นฉบับมีการได้หลากหลายและมีความละเอียดมากขึ้นข้อดีของวิธีการนี้คือ ผู้พัฒนาไม่จำเป็นต้องมีความรู้ทางด้านภาษาศาสตร์ของทั้งสองภาษาจึงทำให้ไม่เกิดปัญหาเรื่องไวยากรณ์ต่างๆ แต่การแปลด้วยวิธีการนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีจำนวนข้อมูลมหาศาลเพื่อหาค่าทางสถิติที่จะนำไปใช้เปรียบเทียบกับประโยคที่จะนำมาแปลได้อย่างครอบคลุม

2.2.3 การแปลภาษาด้วยเครื่องแบบอ้างอิงตัวอย่าง (Example-based Machine Translation)เป็นการแปลภาษาด้วยเครื่องโดยมีฐานข้อมูลคู่ภาษาเป็นองค์ประกอบสำคัญ ซึ่งทำหน้าที่เก็บคำและประโยคตัวอย่างของคู่ภาษาที่ใช้จริงในชีวิตประจำวัน เพื่อนำไปคำนวณหาว่าประโยคที่รับเข้ามานั้น ควรทำการแปลออกมาเป็นประโยคในรูปแบบใด ระบบการทำงานของวิธีการนี้ แบ่งเป็น 2 ส่วนหลัก คือ การสร้างต้นแบบการแปล และการรวมประโยคคำแปลใหม่ โดยในส่วนแรกจะเป็นการจับคู่คำและประโยคตัวอย่างในฐานข้อมูลคู่ภาษา สร้างเป็นต้นแบบการแปลพื้นฐานเพื่อนำไปใช้เปรียบเทียบกับประโยคที่รับเข้ามา จากนั้นในส่วนที่ 2 จะทำการรวบรวมคำแปลที่ได้จากการเปรียบเทียบ มาสร้างเป็นประโยคผลลัพธ์ข้อดีของวิธีการนี้คือ ไม่ใช้กฎไวยากรณ์และพจนานุกรมคู่ภาษา จึงไม่เกิดปัญหาเรื่องไวยากรณ์ต่างๆ อีกทั้งยังช่วยประหยัดเวลาในการแก้ไขปรับปรุงระบบ เพราะใช้เวลาในการพัฒนาระบบที่จะทำงานเชื่อมต่อกับฐานข้อมูลคู่ภาษาเพียงครั้งเดียว ถ้าพัฒนาระบบให้ทำงานได้อย่างมีประสิทธิภาพแล้ว สามารถนำไปใช้กับฐานข้อมูลคู่ภาษาอื่นๆ ได้อีกด้วย แต่การแปลด้วยวิธีการนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีจำนวนคำและประโยคตัวอย่างเป็นจำนวนมาก และต้องทำการจับคู่คำและประโยคตัวอย่างให้มีความถูกต้องแม่นยำ เพื่อให้ได้การแปลที่มีประสิทธิภาพในงานวิจัยนี้จะเลือกใช้การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ ซึ่งเป็นวิธีเริ่มต้นที่มักจะใช้ในการแปลคู่ภาษาใหม่ๆ นอกจากนั้นการแปลภาษาด้วยเครื่องแบบใช้ฐานกฎไม่จำเป็นต้องใช้ฐานข้อมูลคู่ภาษาขนาดใหญ่ ซึ่งทำให้ง่ายต่อการพัฒนาเพื่อแปลประโยคพื้นฐานของคู่ภาษานั้นๆ

**2.2 งานวิจัยที่เกี่ยวข้อง**

ศรายุทธ ปลัดกอง ได้เสนอการพัฒนาโปรแกรมการถอดเสียงอักษรไทยเป็นอักษรโรมันตามวิธีการของราชบัณฑิตยสถาน โดยนำคำ วลี หรือประโยค มาแยกคำตามพจนานุกรมเพื่อแบ่งพยางค์ โดยใช้วิธีการจัดกลุ่มตัวอักขระ และนำกลุ่มอักขระไปแยกคำโดยวิธีเปรียบเทียบคำกับพจนานุกรมแบบฟอร์เวิร์ดและแบ็กเวิร์ด จะได้ผลลัพธ์เป็นหน่วยเสียงระดับพยางค์ แล้วจึงนำแต่ละพยางค์ไปถอดสียงเป็นอักษรโรมันตามวิธีการของราชบัณฑิตยสถาน และเพิ่มสัญลักษณ์แทนระดับเสียงเพื่อใช้แทนการออกเสียงวรรณยุกต์เพื่อให้การออกเสียงมีความถูกต้องมากขึ้น ผลการทดลองพบว่ามีความถูกต้อง 99.95 %จากคำในพจนานุกรมจำนวน 21,399 คำ

อัจฉราภรณ์ ธาตุวิสัย ได้เสนองานการศึกษาวิเคราะห์คีตโควินทกาวยะ โดยมีวัตถุประสงค์ คือ 1. เพื่อปริวรรตและแปลคีตโควินทกาวยะเป็นภาษาไทย 2. เพื่อศึกษาวิเคราะห์เนื้อหาของคีตโควินทกาวยะ และ 3 .เพื่อศึกษาวิเคราะห์วรรณศิลป์ของคีตโควินทกาวยะ ต้นฉบับที่ใช้ในการศึกษาครั้งนี้คือต้นฉบับภาษาสันสกฤต ตัวอักษรเทวนาครีจากหนังสือ คีตโควินทะของชยเทพ ตีพิมพ์โดย สัตสันฆะ เสวะ สะมิถิ ไม่ระบุปี ที,พิมพ์ การวิจัยแบ่งเป็น 2 ส่วน คือ การวิเคราะห์เนื้อหาใน 3 หัวข้อ ได้แก่ การร้อยเรียงเรื่อง องค์ประกอบของเรื่องและแนวคิดของเรื่องและการวิเคราะห์วรรณศิลป์ โดยใช้ทฤษฎีอลังการศาสตร์ 2 ทฤษฎี ได้แก่ทฤษฎีอลังการ และทฤษฎีรส โดยมีขั้นตอนการศึกษา คือ ค้นคว้าและรวบรวมข้อมูลต่างๆ ที่เกี่ยวข้อง ปริวรรตคีตโควินทกาวยะจากอักษรเทวนาครีเป็นอักษรไทย และแปลจากภาษาสันสกฤตเป็นภาษาไทย จากนั้นจึงศึกษาวิเคราะห์ในด้านเนื้อหาและด้านวรรณศิลป์ แล้วอภิปรายและสรุปผลการศึกษา

ชุลีกร กิตติกูล ได้เสนอการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน การถอดคำแบบถ่ายเสียงสำหรับแต่ละคำสามารถสร้างได้จากกฎ หรือใช้แบบจำลอง ทางสถิติ หรือค้นจากพจนานุกรม อย่างไรก็ตามการขาดมาตรฐานและความหลายหลากของการแปลงชื่อบุคคลไทยให้เป็นชื่อ ที่เขียนด้วยอักษรโรมันเป็นงานที่ท้าทาย และแม้ว่าวิธีที่ใช้พจนานุกรมเหมือนจะให้ผลที่ค่อนข้างถูกต้องที่สุด แต่ส่วนของการแปลงตัวอักษรเป็นเสียงก็ยังมีความจำเป็นสำหรับคำที่ไม่พบในพจนานุกรม งานวิจัยนี้เสนอวิธีการถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมันให้เป็นเสียงภาษาไทย โดยคำนึงถึงความนิยมในการใช้งาน ชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมันจะถูกแบ่งให้เป็นสายลำดับของแกรมโดยใช้พจนานุกรมแกรมสะสมซึ่งถูกสร้างจากชื่อมากกว่า 130,000 ชื่อ ผลการศึกษาพบว่าวิธีนี้ให้ความถูกต้องของ

**2.2.5 Syllable-based Thai-English Machine Transliteration**

ชัย วุฒิวิวัฒน์ชัย และอัษฎางค์ แตงไทย

**2.2.5 Machine Transliteration for Indian Languages: A Literature Survey**