

การพัฒนาคลังศัพท์สำหรับการแปลงคำปริวรรตของอักษร ไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยคอมพิวเตอร์

Developing A Lexicon for Computerized Transformation of The Transliterated Words of Thai Noi Characters into Isan Language Pronunciations

จินตนา พลศรี^{1*}, นิสัชล ชำนงศรี¹ และ จีรยุทธ ไชยจารุวนิช²

Jintana Polsri^{1*}, Nisachol Chamnongsri¹, Jeerayut Chaijaruvanit²

¹สำนักวิทยาศาสตร์และศิลปดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี ประเทศไทย

²ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ประเทศไทย

¹Institute of Digital Arts and Science, Suranaree University of Technology, Thailand

²Department of Computer Science, Faculty of Science, Chiang Mai University, Thailand

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาคลังศัพท์สำหรับแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยคอมพิวเตอร์ และประเมินประสิทธิภาพของคลังศัพท์ที่พัฒนาขึ้น แบ่งการวิจัยเป็น 3 ขั้นตอน คือ (1) การศึกษาลักษณะคลังศัพท์ภาษาไทยถิ่นอีสาน 7 เล่ม และพจนานุกรมภาษาไทยมาตรฐาน 1 เล่ม (2) การออกแบบและพัฒนาคลังศัพท์ และ (3) การประเมินประสิทธิภาพของคลังศัพท์ในการนำไปใช้แปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสาน โดยใช้คำปริวรรตของตำรายาโบราณอีสาน 7,929 คำ เป็นข้อมูลชุดทดสอบ และการประเมินประสิทธิภาพด้านความเหมาะสมในการออกแบบและพัฒนาคลังศัพท์ โดยผู้เชี่ยวชาญอักษรโบราณ 12 คน และผู้ใช้ประโยชน์จากตำรายาโบราณอีสาน 27 คน ผลการวิจัยพบว่า คลังศัพท์ภาษาไทยถิ่นอีสานมีข้อจำกัดในการนำไปใช้แปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยคอมพิวเตอร์ เนื่องจากคำศัพท์หลักไม่เป็นแบบคำปริวรรต มีเกณฑ์การสะกดคำอ่านที่หลากหลาย และแบบอักษรไทยน้อยที่จะใช้เป็นคำศัพท์หลักยังไม่มีรหัสยูนิโคด ฐานข้อมูลคลังศัพท์ที่พัฒนาขึ้นมีคำศัพท์หลักเป็นคำปริวรรต 4,645 คำ ชนิดของคำ 36 ชนิด และเกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสาน 10 ข้อ ประสิทธิภาพของคลังศัพท์ในการแปลงคำปริวรรตของตำรายาโบราณอีสานเป็นคำอ่านภาษาไทยถิ่นอีสานมีความถูกต้องร้อยละ 71.02 ความเหมาะสมในการออกแบบและพัฒนาฐานข้อมูลคลังศัพท์มีค่าเฉลี่ยทุกด้านอยู่ในระดับดี ($\bar{X} = 3.89$, $SD = 0.71$) การใช้ศัพท์หลักเป็นคำปริวรรตจะช่วยลดการสร้างคลังคำศัพท์เฉพาะของอักษรโบราณแต่ละอักษรได้

คำสำคัญ: คลังศัพท์, การปริวรรต, การปริวรรตภาษาโบราณด้วยเครื่อง, การแปลภาษาด้วยเครื่อง

ABSTRACT

The objectives of this research were to design and develop the lexicon for the transformation of the transliterated words of Thai Noi characters into Isan language pronunciations by computer and to evaluate the efficiency of the lexicon being developed. The research methodology was separated into 3 steps: 1) the study of the characteristics of Isan lexicons totally from 7 books, and one standard Thai lexicon, 2) the design and development of the lexicon and 3) the evaluation of the efficiency of the lexicon for transformation of the 7,929 transliterated words as a testing data into Isan language pronunciation and the evaluation of the efficacy of the design and development of the lexicon by 12 ancient linguists and 27 users of the ancient Isan medicine textbooks. The results of the research revealed that the Isan lexicons were limited for the transformation of the transliterated words of Thai Noi characters into Isan language pronunciations by computer. This is because main vocabularies were not the transliterated words of Thai Noi characters and had a variety of spelling criteria. And Thai Noi font that will be used as main words do not have a Unicode code. The designed and developed lexicon contained 4,645 main vocabularies, 36 parts of speech, 10 spelling criterias of Isan language pronunciation. The result of the accuracy of transformation of the transliterated ancient Isan medicine recipes into Isan language pronunciation was 71.02 percent and the overall mean of the design and development of the lexicon efficacy was in a good level for all items ($\bar{X} = 3.89$, $SD = 0.71$). Using the main vocabulary as a transliteration reduces the need to build a unique vocabulary for each ancient script.

KEYWORDS: Lexicon, transliteration, machine transliteration of ancient language, machine transliteration

*Corresponding Author: jpolsri@yahoo.com¹, nisachol@sut.ac.th², jeerayut.c@cmu.ac.th³

Received: 11/07/2022; Revised: 20/02/2023; Accepted: 31/03/2023

1. บทนำ

เอกสารโบราณเป็นหลักฐานชั้นต้นที่มีความสำคัญต่อการศึกษาร่องรอยชีวิตความเป็นอยู่ และองค์ความรู้ของผู้คนในอดีตในหลากหลายด้าน อาทิ ประวัติศาสตร์ อักษรศาสตร์ กฎหมายโบราณ และตำรายา เป็นต้น อย่างไรก็ตาม การเข้าถึงและนำองค์ความรู้เหล่านั้นมาใช้สำหรับคนทั่วไปยังเป็นเรื่องที่ยาก เนื่องจากเอกสารโบราณบันทึกด้วยอักษรและภาษาโบราณที่ปัจจุบันเลิกใช้ไปแล้ว ดังนั้นเพื่อให้คนที่ไม่มีความรู้ด้านอักษรและภาษาโบราณสามารถนำไปศึกษาได้ง่ายขึ้นจำเป็นต้องทำการปริวรรต (Transliterations) เป็นอักษรและภาษาที่คนปัจจุบันสามารถอ่านได้ โดยการปริวรรตเอกสารโบราณเป็นกระบวนการถอดอักษรโบราณไปเป็นอักษรและอักษรวิธีการเขียนแบบปัจจุบันอย่างมีหลักการ เมื่อปริวรรตแล้วสามารถเขียนกลับเป็นอักษรและอักษรวิธีการเขียนแบบเดิมได้ (Wimonkasem, 2005)

การปริวรรตอักษรโบราณเป็นอักษรไทยปัจจุบันมี 3 รูปแบบ คือ (1) ถอดรูปอักษรตรงตัว ตัวต่อตัวตามต้นฉบับโดยไม่คำนึงถึงเสียงอ่าน นิยมใช้กับการปริวรรตจารึก (2) ถอดรูปอักษรและสะกดคำตรงตามเสียงในภาษาที่เขียน มักใช้กับวรรณกรรมท้องถิ่น และ (3) ถอดรูปอักษรและสะกดคำตามอักษรวิธีของภาษาไทยปัจจุบัน โดยไม่คำนึงถึงเสียงในภาษาถิ่นและอักษรวิธีการเขียนในเอกสารต้นฉบับเพื่อให้อ่านเข้าใจได้ง่ายขึ้น (Na Nagara, 1998) การปริวรรตแบ่งเป็น 2 ขั้นตอน ได้แก่ (1) การทำ “คำปริวรรต” คือ การถอดรูปอักษรโบราณเป็นอักษรไทยปัจจุบันแบบตรงตัวกับเอกสารต้นฉบับ และ (2) การทำ “คำอ่าน” คือ การทำปริวรรตเป็นคำอ่านแบบภาษาถิ่น หรือคำอ่านแบบภาษาไทยมาตรฐานที่สะกดคำตามพจนานุกรม ส่วนคำที่เป็นชื่อเฉพาะของบุคคล ชื่อสถานที่ ชื่อเมือง จะคงรูปแบบการสะกดตามรูปแบบเดิม และเขียนเชิงอรรถอธิบายคำที่ใช้ในปัจจุบันเอาไว้เพื่อให้รู้จักคำดั้งเดิม สำหรับวรรณกรรมที่มีวิธีเขียนใกล้เคียงกับอักษรวิธีภาษาไทยปัจจุบันค่อนข้างมากอาจตัดขั้นตอนการทำคำปริวรรตออกไป (Paiboonwangcharoen, 1999)

อย่างไรก็ดี การปริวรรตเอกสารโบราณโดยมนุษย์เป็นกระบวนการที่ต้องใช้เวลาและมีค่าใช้จ่ายสูง เนื่องจากต้องอาศัยผู้ปริวรรตที่มีประสบการณ์สูงและมีความเชี่ยวชาญด้านการอ่านอักษรและภาษาโบราณเป็นผู้ดำเนินการ ซึ่งปัจจุบันบุคลากรด้านนี้มีจำนวนน้อยมากเมื่อเทียบกับปริมาณเอกสารโบราณมีอยู่เป็นจำนวนมากมหาศาล ความ

พยายามในการนำเทคโนโลยีมาช่วยให้การปริวรรตเอกสารโบราณทำได้รวดเร็วยิ่งขึ้น อาทิ การเพิ่มความชัดเจนของตัวอักษรให้อ่านได้ง่ายขึ้น (Bianco *et.al.*, 2010; Hollaus, 2012; Natarajan & Sreedevi, 2017) การแปลงภาพอักษรโบราณเป็นอักษรโบราณอิเล็กทรอนิกส์เพื่อนำไปประมวลผลด้วยคอมพิวเตอร์ได้ (Romulus, 2010; Chamchong *et al.*, 2011; Chueaphun *et al.*, 2012; Inkeaw, 2017; Phattarachairavee, 2019) รวมถึงงานด้านการปริวรรตและแปลภาษาไทยโบราณด้วยเครื่อง เช่น การแปลงอักษรธรรมอีสานเป็นคำอ่านภาษาไทยถิ่นอีสาน (Paiboon, 2009) และภาษาไทยมาตรฐาน (Phromsoda, 2010) การวิเคราะห์โครงสร้างประโยค (Suradet, 2011) และการพัฒนาระบบการแบ่งคำอักษรธรรมอีสาน (Somsap & Seresangtakul, 2020) เป็นต้น

การแปลภาษาด้วยเครื่อง (Machine Translation) เป็นการประยุกต์ใช้เทคโนโลยีด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ที่ทำให้คอมพิวเตอร์เข้าใจภาษาที่มนุษย์ใช้สื่อสารกันทั้งในรูปแบบภาษาพูดและภาษาเขียน ช่วยลดข้อจำกัดในการสื่อสารภาษาที่แตกต่างกันและลดต้นทุนการแปลภาษา กระบวนการแปลภาษาด้วยเครื่องจะแตกต่างกันไปตามลักษณะเฉพาะของแต่ละภาษาและในหลายกระบวนการต้องอาศัยคลังศัพท์ (Lexicon) ในการประมวลผล เช่น การตัดคำ (Tokenization) การแบ่งส่วนหรือการกำกับชนิดของคำ (Part of Speech: POS) โดยเฉพาะการทำข้อความให้เป็นบรรทัดฐาน (Text normalization) จะใช้คลังศัพท์เพื่อปรับเปลี่ยนคำที่มีความหมายเดียวกันแต่เขียนได้หลายแบบมาเขียนให้อยู่ในรูปแบบเดียวกันด้วยวิธี เช่น การตรวจแก้ตัวสะกด (Spelling Correction) การลดรูปคำศัพท์ (Stemming) การผันคำ (Inflection) การสะกดคำตามรากศัพท์ (Root) หรือใช้คำตามพจนานุกรม (Lemmatization) เป็นต้น รวมทั้งกระบวนการแปลภาษาที่ใช้คลังศัพท์เป็นฐานในการแปลภาษา ทั้งนี้ประสิทธิภาพการแปลภาษาด้วยเครื่องที่ใช้คลังศัพท์เป็นฐานในการแปลจะขึ้นอยู่กับจำนวนคำศัพท์ในคลังศัพท์เป็นสำคัญ (Azizan *et al.*, 2019; Intasorn., 2021) เช่นเดียวกับงานด้านการปริวรรตและการแปลภาษาไทยโบราณด้วยเครื่อง (Paiboon, 2009; Phromsoda, 2010; Suradet, 2011; 2554; Somsap & Seresangtakul, 2020) ส่วนใหญ่จะใช้คลังศัพท์เป็นฐานในแปล การตัดคำ และการกำกับชนิดของคำ (Part-Of-Speech: POS) โดยโครงสร้างหลักของการ

อธิบายคำศัพท์ในคลังศัพท์ภาษาไทยโบราณที่พบประกอบด้วย คำศัพท์หลักอักษรไทยโบราณ (Ancient Thai font) และคำศัพท์หลักอักษรไทยปัจจุบันภาษาไทยถิ่น หรือภาษาไทยมาตรฐาน ชนิดของคำ และความหมายของคำ

คลังศัพท์ (Lexicon) หรือ ศัพท์านุกรม เป็นศัพท์บัญญัติ มีชื่อเรียกอื่น ๆ เช่น พจนานุกรม ปทานุกรม หรืออภิธานคำศัพท์ หมายถึง หนังสือสำหรับค้นคว้าความหมายของคำที่เรียงตามลำดับบท (Office of the Royal Society, n.d.) การอธิบายคำศัพท์จะประกอบด้วย คำศัพท์หลัก (Terms) ความหมายของคำศัพท์ (Definition) สัทอักษร (Phonetics) ไวยากรณ์ (Syntax) และโครงสร้างของหน่วยคำในประโยคหรือสัทฐานวิทยา (Morphology) เช่น คำนาม คำกริยา เป็นต้น ทั้งนี้ขึ้นอยู่กับวัตถุประสงค์ของการสร้างคลังศัพท์ (Hirst, 2009) การสร้างคลังศัพท์จำแนกตามแหล่งข้อมูลที่น่าสนใจคัดเลือกคำศัพท์ได้ 2 วิธี คือ (1) สร้างจากพจนานุกรมที่มีอยู่แล้ว (Dictionary-based Approach) ตัวอย่าง เช่น คลังศัพท์เครือข่ายภาษาไทย (Thai WordNet) สร้างจากพจนานุกรมภาษาไทย-อังกฤษและ WordNet (Thoongsup, 2009) คลังศัพท์บอกความรู้สึกภาษาอังกฤษสร้างจาก WordNet (Eduard., 2010) คลังศัพท์บอกความรู้สึกภาษาไทย สร้างจากพจนานุกรมภาษาไทย (Intasorn *et al.*, 2021) ฐานข้อมูลพจนานุกรมคำร่วมเชื้อสาย สร้างจาก Wang Li's Dictionary of Cognate Words (Fang & Xu, 2022) เป็นต้น (2) สร้างจากคลังข้อมูลภาษา (Corpus-bases Approach) ซึ่งเป็นคลังภาษาเขียนหรือภาษาพูดที่ใช้อยู่จริงที่ได้มีการรวบรวมไว้แล้ว หรือรวบรวมขึ้นมาใหม่จากแหล่งต่าง ๆ ตัวอย่างเช่น พจนานุกรม Longman Dictionary of American English สร้างจาก Longman Corpus Network ขนาด 330 ล้านคำ พจนานุกรม Macmillan สร้างจาก World English Corpus ขนาด 200 ล้านคำ คลังศัพท์บอกความรู้สึกในภาษาไทยจากบทวิจารณ์ออนไลน์ สร้างจากคลังข้อมูลภาษาที่รวบรวมขึ้นมาใหม่จากบทวิจารณ์โรงแรม ภาพยนต์ และบทวิจารณ์แอปพลิเคชันบนโทรศัพท์มือถือ 4,736 บทวิจารณ์ (Lorattanachaiyong, 2017) คลังศัพท์วิชาการทางภาษาอังกฤษธุรกิจ สร้างจาก KRU-BE Corpus ที่รวบรวมคำจากตำรา บทความวิจัย บทความออนไลน์ และหนังสือพิมพ์สาขาวิชาภาษาอังกฤษธุรกิจ ขนาด 16 ล้านคำ (Tongpoon-Patanasorn, 2018) พจนานุกรมคำศัพท์แสดงความเกลียดชัง สร้างจากการรวบรวมความคิดเห็นที่มีต่อข่าวใน Line Today จำนวน

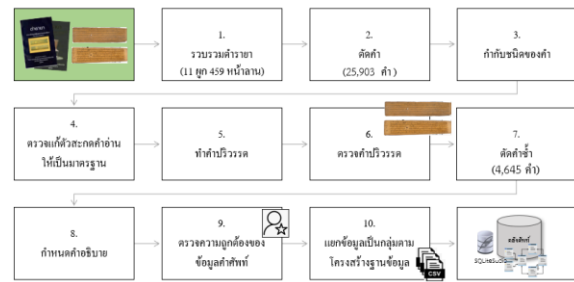
11,917 ความคิดเห็น (Wang, Day, & Wu, 2022) เป็นต้น

อักษรไทยน้อยเป็นอักษรที่คนไทยถิ่นอีสานและคนลาวที่อาศัยในอาณาจักรล้านช้าง (ปัจจุบันคือบริเวณแถบลุ่มแม่น้ำโขงและสปป.ลาว) ใช้จดบันทึกตั้งแต่วราช.ศ. 1890 เป็นต้นมา มักใช้บันทึกเนื้อหาทางโลก อาทิวรรณกรรมและนิทานพื้นบ้าน ตำรายา โบราณศาสตร์กฎหมาย พงศาวดาร และเอกสารทางราชการ เป็นต้น โดยบันทึกเป็นภาษาไทยถิ่นอีสานหรือภาษาลาว ส่วนมากพบในเอกสารโบราณ ประกอบด้วยพยัญชนะเดี่ยวประมาณ 24 รูป พยัญชนะประสม (พยัญชนะ 2 ตัวเขียนติดกัน) 8 รูป สระ 23 รูป ไม่มีรูปวรรณยุกต์ อักษรวิธีหรือวิธีประสมอักษรให้เป็นคำส่วนใหญ่จะคล้ายคลึงกับอักษรไทยปัจจุบัน คือ วางพยัญชนะต้นและตัวสะกดไว้บนบรรทัด เขียนสระไว้รอบพยัญชนะต้น (หน้า หลัง บน ล่าง) ที่แตกต่างคือ มีการใช้รูปพยัญชนะ ตัวเชิง หรือ ตัวเพื่อง ของอักษรธรรมซึ่งเป็นอักษรโบราณอีกชนิดหนึ่งที่คนไทยถิ่นอีสานนิยมใช้บันทึกเนื้อหาทางด้านพุทธศาสนาหรือคาถาอาคม มาเขียนร่วมด้วย (Punnotok, 2006) ซึ่งมักพบเขียนปะปนกันอยู่เสมอ (Kaewplang, 2013) อักษรไทยน้อยเลิกใช้ไปเมื่อมีการจัดการเรียนการสอนในระบบโรงเรียนตามพระราชบัญญัติประถมศึกษา พ.ศ. 2465 ซึ่งมีนโยบายให้คนไทยทุกคนเรียนอักษรและภาษาไทยมาตรฐานเหมือนกันทั่วประเทศ (Punnotok, 2006)

จากการสัมภาษณ์นักภาษาโบราณและการวิเคราะห์การเขียนอักษรไทยน้อยในตำรายาโบราณอีสาน (Poltri, 2021) พบว่า ปัญหาการอ่านและปริวรรตอักษรไทยน้อยที่จารในใบลานด้วยภาษาอีสาน ได้แก่ (1) ความหลากหลายของรูปอักษรจากลายมือเขียนที่มีลีลาแตกต่างกันของผู้เขียนแต่ละคน ลายมือที่อ่านยากจะส่งผลต่อความถูกต้องของการระบุรูปอักษรแต่ละตัว (2) การสร้างคำโดยใช้วิธีประสมอักษรด้วยอักษรวิธีพิเศษเฉพาะบุคคลของผู้เขียนที่แตกต่างไปจากวิธีประสมอักษรปกติ เช่น การใช้พยัญชนะตัวสะกดของคำพยางค์หน้าเป็นพยัญชนะต้นของคำพยางค์หลัง (ใช้พยัญชนะร่วมกัน) การสร้างกฎเกณฑ์วิธีประสมอักษรขึ้นมาใหม่โดยเลียนแบบวิธีประสมอักษรที่ใช้นักภาษาบาลี หรือ การใช้สัญลักษณ์แทนคำหรือตัวอักษร ซึ่งพบในงานของ สมจิตศรีปัญญา (Somjitsripanya, 1982) จันทร์แดง คำลือหาญ (Khamluehan, 1992) กิตติวัฒน์ สัตนาโค (Sattanakho, 1999) และ เอ็ม อร ชาวน์สวน (Chawsuan, 2015) เช่นกัน และ (3) อักษรไทยน้อยมี

ลักษณะ “เขียนอย่างหนึ่งแต่อ่านอีกอย่างหนึ่ง” คือ คำหนึ่งคำสามารถอ่านได้หลายแบบ ซึ่งการอ่านแต่ละแบบจะทำให้คำมีความหมายต่างกัน เช่น “ไก” (คำปริวรรต) อ่านได้ 3 แบบ ได้แก่ “ไก” “ไกล” “ไกล้” หรือ “ป้า” (คำปริวรรต) อ่านได้ 3 แบบ ได้แก่ “ป้า” “ป้า” “ปลา” หรือกรณีคำที่อ่านออกเสียงเหมือนกันแต่สามารถสะกดได้หลายแบบ เช่น “ไซ” พบรูปแบบการสะกดทั้งแบบใช้พยัญชนะต้น “ค” และ “ช” คือ “ไซ” “ไค” ผู้อ่านจะต้องผันเสียงให้ตรงกับความหมายตามบริบทของเรื่องที่อ่านจึงจะอ่านได้ถูกต้อง ซึ่งสาเหตุเกิดจากอักษรไทยน้อยมีจำนวนตัวอักษรน้อยกว่าอักษรไทยปัจจุบัน ไม่มีวรรณยุกต์กำกับโทนเสียง และการสะกดคำมุ่งเน้นการอ่านให้ตรงกับเสียงพูดจึงไม่เคร่งครัดเรื่องตัวสะกด ดังนั้น ผู้ปริวรรตจึงจำเป็นต้องรู้ภาษาไทยถิ่นอีสานโบราณ และมีความรู้เกี่ยวกับเรื่องที่จะอ่านเป็นอย่างดี จึงจะทำคำอ่านได้ถูกต้อง

การปริวรรตเอกสารโบราณที่บันทึกด้วยอักษรไทยน้อยเป็นคำอ่านอักษรไทยปัจจุบัน ทั้งคำอ่านภาษาไทยถิ่นอีสานและคำอ่านภาษาไทยมาตรฐานแบบอัตโนมัติด้วยเครื่องนั้น จำเป็นต้องใช้คลังศัพท์ในขั้นตอนการตัดคำ (Tokenization) และการทำคำอ่าน ซึ่งจากสำรวจเบื้องต้นพบว่า คลังศัพท์ที่มีคำศัพท์หลักพิมพ์ด้วยอักษรไทยน้อยมีเพียง 1 เล่ม (Chantao, 2015) และยังไม่สามารถสร้างเป็นพจนานุกรมอิเล็กทรอนิกส์ เนื่องจาก ปัจจุบันแบบอักษรไทยน้อย (Thai Noi font) ที่จะใช้พิมพ์คำศัพท์หลักยังไม่มีรหัสยูนิโคด (Unicode) เพื่อแก้ไขปัญหาดังกล่าวงานวิจัยนี้จึงนำเสนอการพัฒนาคลังศัพท์อิเล็กทรอนิกส์ที่มีคำศัพท์หลักเป็นคำปริวรรตของอักษรไทยน้อย (คำถ่ายถอดจากรูปอักษรไทยน้อยเป็นอักษรไทยปัจจุบันแบบตรงตัว) แทนคำศัพท์หลักอักษรไทยน้อย ในหมวดตำรายาโบราณอีสาน เพื่อช่วยให้การอ่านและปริวรรตตำรายาโบราณอีสานทำได้รวดเร็วขึ้น เป็นประโยชน์ต่อการนำไปศึกษาค้นคว้าต่อไป โดยคลังศัพท์ที่พัฒนาขึ้นจะช่วยให้กระบวนการตัดคำและแปลงคำปริวรรตของระบบปริวรรตและแปลตำรายาโบราณอีสานที่บันทึกด้วยอักษรไทยน้อยด้วยคอมพิวเตอร์ทำได้ถูกต้องและเร็วยิ่งขึ้น นอกจากนี้ยังสามารถนำไปใช้ในการประมวลผลภาษาธรรมชาติ (การตัดคำ การกำกับชนิดของคำ การแปล) การรู้จำชื่อเฉพาะ (Name Entity Recognition) เช่น ชื่อโรค ชื่อเรียกอาการ และชื่อสมุนไพร (พืชวัตถุ สัตว์วัตถุ ธาตุวัตถุ) การสกัดและสรุปสารสนเทศในตำรายาโบราณอีสานเพื่อพัฒนาระบบจัดทำดัชนีและจัดหมวดหมู่ตำรายาโบราณแบบอัตโนมัติ และระบบสืบค้นตำรายาโบราณ



รูปที่ 1 ขั้นตอนการรวบรวมคำศัพท์

ต่อไป อีกทั้งยังช่วยปกป้องและสงวนรักษามรดกภูมิปัญญาของชาติด้านตำรายาและด้านภาษาโบราณซึ่งเป็นมรดกทางวัฒนธรรมที่จับต้องไม่ได้ (Intangible Cultural Heritage) ให้คงอยู่และสามารถต่อยอดนวัตกรรมได้ในอนาคต การวิจัยนี้ เป็นส่วนหนึ่งของการวิจัยเรื่องการพัฒนากรอบการทำงานในการแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสาน (Polstri, 2021) และได้ผ่านการรับรองจริยธรรมการวิจัยในมนุษย์ เอกสารรับรองเลขที่ COA No. 64/2563

2. วิธีดำเนินการวิจัย

แบ่งการวิจัยเป็น 3 ขั้นตอน ได้แก่ การวิเคราะห์ลักษณะคลังศัพท์ภาษาไทยถิ่นอีสาน การออกแบบและพัฒนาคลังศัพท์ และการประเมินประสิทธิภาพคลังศัพท์ดังนี้

2.1 การวิเคราะห์ลักษณะคลังศัพท์ภาษาไทยถิ่นอีสาน

เพื่อให้ทราบว่าคลังศัพท์ภาษาไทยถิ่นอีสานที่มีอยู่ในปัจจุบันมีลักษณะอย่างไร และสามารถนำมาใช้ในกระบวนการแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานได้หรือไม่ กลุ่มตัวอย่างในการศึกษา ได้แก่ คลังศัพท์จำนวน 8 เล่ม ดังนี้ คลังศัพท์ภาษาไทยมาตรฐาน 1 เล่ม คือ พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 (Office of the Royal Society, 2011) และคลังศัพท์ภาษาไทยถิ่นอีสาน จำนวน 7 เล่ม ดังนี้ สารานุกรมภาษาอีสาน-ไทย-อังกฤษ (Pinthong, 1989) ประมวลคำศัพท์ภาษาลาวในเอกสารโบราณ เล่ม 1 - 5 (National Library of Thailand, 2002) พจนานุกรม 3 ภาษา (ไทยน้อย-อังกฤษ-ไทย) (Chantao, 2015) พจนานุกรมภาคอีสาน-ภาคกลาง ฉบับปณิธานสมเด็จพระมหาธีรราชเจ้า (ดิสสนมหาเถระ) (Somdet Phra Maha Wirawong (Tissamah Thera), 1998) พจนานุกรมภาษาถิ่นอีสาน: เว้าอีสาน (Phimworamethakun & Phimworamethakun, 2002) พจนานุกรมภาษาอีสานฉบับ คำพูน บุญทวี (Suksang, 2005)

และ พจนานุกรมภาษาอีสาน-ไทยกลาง (Raksuthi, 2011) รวบรวมข้อมูลโดยใช้แบบวิเคราะห์ลักษณะคลังคำศัพท์ภาษาไทยถิ่นอีสาน ที่ได้จากการวิเคราะห์ลักษณะคลังคำศัพท์ที่ใช้ในการปริวรรตเอกสารโบราณ 11 รายการ ได้แก่ แบบอักษรที่ใช้พิมพ์คำศัพท์หลัก ภาษาของคำศัพท์หลัก หมวดคำศัพท์ มีคำอ่านภาษาไทยถิ่นอีสานโบราณ มีคำอ่านภาษาไทยมาตรฐาน มีชนิดของคำ (POS) มีคำอธิบายที่มาของอธิบาย มีตัวอย่างประโยคมีคำตรงข้าม มีคำคล้ายคลึง และมีคำปริวรรต

2.2 การออกแบบและพัฒนาคลังคำศัพท์

การออกแบบและพัฒนาคลังคำศัพท์แบ่งเป็น 3 ขั้นตอนหลักดังนี้

2.2.1 การออกแบบและสร้างฐานข้อมูลคลังคำศัพท์ มี 4 ขั้นตอน ได้แก่ (1) การออกแบบโครงสร้างฐานข้อมูลคลังคำศัพท์ในระดับแนวคิด (Conceptual Schema Design) โดยการวิเคราะห์เอนทิตี (Entity) แอททริบิวต์ (Attribute) ของแต่ละเอนทิตี และความสัมพันธ์ระหว่างเอนทิตี (Relationship) และในระดับกายภาพ (Physical Design) ได้แก่ การเลือกระบบฐานข้อมูลที่จะใช้สร้างและจัดเก็บคลังคำศัพท์ (2) การสร้างฐานข้อมูลคลังคำศัพท์ตามรูปแบบไว้ด้วยโปรแกรมจัดการฐานข้อมูล SQLite Studio (3) การวิเคราะห์เกณฑ์การสะกดคำศัพท์ในคลังคำศัพท์ภาษาไทยถิ่นอีสาน (เอกสารชุดเดียวกันกับที่ใช้วิเคราะห์ลักษณะคลังคำศัพท์ภาษาไทยถิ่นอีสาน) ร่วมกับลักษณะของภาษาไทยถิ่นอีสาน (Kingkham, 2013) เพื่อสร้างเกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสาน และ (4) กำหนดชนิดของคำศัพท์ (POS)

2.2.2 การรวบรวมคำศัพท์ ประกอบด้วย 10 ขั้นตอน (รูปที่ 1) ได้แก่ (1) รวบรวมข้อมูลจากตำรายาที่ปริวรรตมาจากตำรายาโบราณอีสานที่บันทึกไว้ในใบลานด้วยอักษรไทยน้อย ได้แก่ ตำรายาวิธมหาชัย จังหวัดมหาสารคาม เล่ม 5 และตำรายาวิธท่าม่วง อำเภอสหัสขันธ์ จังหวัดร้อยเอ็ด เล่ม 2 นำมาคัดเลือกเฉพาะฉบับที่ปริวรรตมาจากอักษรไทยน้อยจำนวน 11 ผูก 459 หน้า ลาน ขนาดคำศัพท์ 25,903 คำ (2) นำตำรายามาตัดคำด้วยมือด้วยโปรแกรมไมโครซอฟต์เวิร์ด โดยรอบแรกใช้เกณฑ์การตัดคำแบบยาวที่สุด และรอบที่ 2 ใช้เกณฑ์ตัดคำแบบสั้นที่สุด สำหรับแยกคำประสมเป็นคำโดด (3) ตรวจสอบการสะกดคำให้เป็นมาตรฐานเดียวกัน (4) กำกับชนิดของคำ (POS) (5) ทำคำปริวรรตของคำศัพท์แต่ละคำ โดยใช้หลักการปริวรรตอักษรไทยน้อยเป็นอักษรไทยปัจจุบันมาเป็นเกณฑ์ในการทำคำปริวรรต (6) ตรวจสอบความถูกต้องโดยเทียบกับสำเนาภาพดิจิทัล

ตำรายาต้นฉบับ (7) ตัดคำซ้ำ โดยนำคำศัพท์ไปจัดเรียงตามลำดับตัวอักษรในโปรแกรมไมโครซอฟต์เอ็กเซล จากนั้นตัดคำศัพท์ที่ซ้ำออก (8) กำหนดรายละเอียดการอธิบายคำศัพท์ตามโครงสร้างฐานข้อมูลคลังคำศัพท์ที่ออกแบบไว้ โดยคำอธิบายความหมายของคำศัพท์ได้จากการสืบค้นจากคลังคำศัพท์ภาษาอีสานชุดเดียวกับคลังคำศัพท์ภาษาไทยถิ่นอีสานที่นำมาศึกษาในข้อ 2.2.1 และสืบค้นเพิ่มเติมจากสารานุกรมรวมหลักเภสัชกรรมไทย (Wuthithamwet, 1997) เว็บไซต์ฐานข้อมูลสมุนไพร คณะเภสัชศาสตร์ มหาวิทยาลัยอุบลราชธานี (www.phargarden.com) และเว็บไซต์เมดไทย (www.medthai.com) โดยคำศัพท์ที่ไม่สามารถค้นหาคำอธิบายได้จะใช้วิธีสอบถามจากนักภาษาโบราณ หรือหมอยาพื้นบ้านอีสาน และระบุแหล่งสืบค้นข้อมูลเป็น “สัมภาษณ์” (9) ตรวจสอบความถูกต้องของคำศัพท์โดยนักภาษาโบราณซึ่งเป็นผู้ปริวรรตตำรายาฉบับที่นำมาศึกษา และ (10) แยกข้อมูลจัดกลุ่มข้อมูลตามโครงสร้างฐานข้อมูลที่ออกแบบไว้ บันทึกข้อมูลแบบไฟล์ CSV และนำเข้า (Import) จัดเก็บในฐานข้อมูล

การพัฒนาคลังคำศัพท์ครั้งนี้ใช้วิธีตัดคำด้วยมือ เนื่องจากมีข้อจำกัดที่สำคัญต่อการตัดคำแบบอัตโนมัติคือ ปัจจุบันยังไม่มีฐานความรู้ที่จะนำมาใช้ในการตัดคำ เช่น พจนานุกรมอิเล็กทรอนิกส์คำศัพท์เฉพาะด้านตำรายาโบราณอีสาน หรือคลังข้อความตำรายาโบราณอีสาน (Corpus) ขนาดใหญ่ สำหรับใช้เป็นข้อมูลฝึกและสร้างโมเดลสำหรับตัดคำแบบอัตโนมัติ อย่างไรก็ตามได้นำแนวคิดการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) และการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) มาใช้ในการตัดคำ ดังนี้

รอบที่ 1 ใช้การเทียบคำที่ยาวที่สุด เนื่องจากคำศัพท์ที่บันทึกในตำรายาโบราณอีสานมีคำประสมที่เป็นคำเฉพาะค่อนข้างมาก โดยเฉพาะชื่อยา/ ชื่อโรค/ ชื่อเรียกอาการ และชื่อสมุนไพร การตัดคำจะดูตามบริบทของตำรายา ได้แก่ องค์ประกอบด้านเนื้อหาที่เขียนในตำรายาอีสานโบราณ 12 กลุ่ม (ชื่อยา/โรค/อาการ ชื่อสมุนไพร สัตส่วน/ปริมาณที่ใช้ วิธีไปเก็บสมุนไพร วิธีเตรียมยา วิธีใช้ยา ผู้ป่วย วันและเวลา คาถา พิธีกรรม คำรักษา และเนื้อหาเรื่องอื่น ๆ) ร่วมกับโครงสร้างการเขียนบรรยายเนื้อหาตำรายา 3 ประเภท (ตำรายาที่ใช้เฉพาะสมุนไพรในการรักษา ตำรายาที่ใช้สมุนไพรร่วมกับคาถา และตำรายาที่ใช้เฉพาะคาถาในการรักษา) 15 รูปแบบ (Polstri, 2021) ตัวอย่างการตัดคำเช่น “ยาแก้ขับเขี้ยวให้เอาเขี้ยวหมาช้างคางกะท่าง...” เป็นตำรายาประเภทใช้

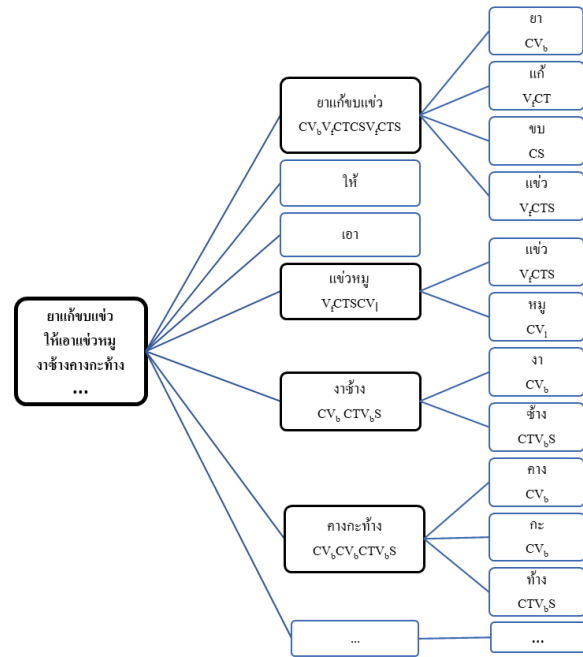
ตารางที่ 1 กฎการตัดคำสำหรับคำที่ขึ้นต้นด้วยสระหน้า

ข้อ	กฎ	ตัวอย่างคำที่ได้ (หลังตัดคำแล้ว)
1	$V_f + C$	ไซ โฟ โบ แล
2	$V_f + C + S$	โคก โคน แดง โทด
3	$V_f + C + V_u + S$	เม็ด เหม็ด เป็ด
4	$V_f + C + T$	แม่ ไซ ไม้ ใหญ่ เหล้า
5	$V_f + C + T + S$	เส้น เป็น เหล็ก แก่น
6	$V_f + C + V_u + T + S$	เบ้ง
7	$V_f + C + V_u + "ย"$	เสีย เมีย เพลีย
8	$V_f + C + V_u + T + "ย"$	เบีย เพีย
9	$V_f + C + V_u + T + "ย" + S$	เอียน เสียบ เหยี่ยว
10	$V_f + C + V_u + "อ"$	เกลือ เครือ
11	$V_f + C + V_u + "อ" + S$	เปลือก เลือด เหลือง
12	$V_f + C + V_u + T + "อ"$	เปือ
13	$V_f + C + V_u + T + "อ" + S$	เอียน (ปลาไหล) เอียน
14	$V_f + C + V_b$	เขา ก่า เภา เป่า เสา
15	$V_f + C + V_b + S$	โพน
16	$V_f + C + T + V_b$	เข้า เป่า โปะ น้ำ
17	$V_f + C + V_u + S$	เกิน

ตารางที่ 2 กฎการตัดคำสำหรับคำที่ขึ้นต้นด้วยพยัญชนะ

ข้อ	กฎ	ตัวอย่างคำที่ได้ (หลังตัดคำแล้ว)
1	$C + T$	บ่ กี่ หน่อ
2	$C + T + C$	ก่น (ชุด) ตัน
3	$C + C$	ขบ ตก ฝน ขน
4	$C + V_b$	ชา คา ต้า งา ยา
5	$C + V_b + S$	หมาก กาม ขาว
6	$C + T + V_b$	ข้า ม่า น้ำ ญ่า
7	$C + T + V_b + S$	บ้าน ข้าว ก้าน
8	$C + \text{ไม่ทันอากาศ} + S$	จัว หนั่ง หัก พัน
9	$C + V_u$	ดี พี
10	$C + V_u + S$	ดิน มิน (ลิมตา) ดิน ผัง
11	$C + V_u + T$	ซึ่ จึ่
12	$C + V_u + T + S$	จิ้น (และ) ดิน ครั้ง ขึ้น
13	$C + V_b + T$	ก่า
14	$C + V_l$	หมู หู
15	$C + V_l + T$	หมู ผู้
16	$C + V_l + T + S$	อุ่น มุ่น (แหลก)
17	$C + V_l + S$	ขุด พุด (ฟัน) ลูก ผูก
18	$C + V_u + "อ"$	มือ
19	$C + "อ"$	คอ
20	$C + "อ" + S$	ออก ดอก ยอด ตอด
21	$C + "อ" + T$	คอ (โรคชนิดหนึ่ง) หน่อ
22	$C + T + "อ" + S$	ด่อน (เผือก) ก้อน น้อย
23	$C + "ว" + S$	กล้วย หน่วย
24	$C + V_u + S + S + G$	จันทน์ จันทร

เฉพาะสมุนไพรมีสมุนไพรมากกว่า 1 ชนิด ตัดคำได้เป็น “ยาแก้ขับแหวให้เอาแหวหมูงาข้างคางกะท่าง...” ซึ่งทำให้กำหนดประเภทของคำ (POS) ในคำประสมที่สร้างจากคำโดดหลายคำทำได้ง่ายขึ้น เช่น “ยาแก้ขับแหว” (ชื่อยา-นาม) “เหลืองตั้งข้าหมื่น” (ลักษณะอาการป่วย-วิเศษณ์) “ฮากดอกเข็ม” (สมุนไพรม-นาม) เป็นต้น



รูปที่ 2 ตัวอย่างวิธีการและผลลัพธ์การตัดคำ

รอบที่ 2 ใช้การเทียบคำที่สั้นที่สุด เพื่อเพิ่มจำนวนคำศัพท์ในคลังศัพท์ให้ได้มากขึ้น การระบุขอบเขตของคำหรือพยางค์จะใช้กฎที่ได้จากการวิเคราะห์การผสมอักษรที่ใช้เขียนตำรายาโบราณอีสาน โดยนำคำที่ผ่านการตัดคำในรอบที่ 1 เฉพาะคำที่มีความยาวตั้งแต่ 5 ตัวอักษร ขึ้นไป (คำศัพท์ที่เป็นคำประสมในตำรายาโบราณอีสานส่วนมากจะมีจำนวนตัวอักษรตั้งแต่ 5 ตัวอักษร) มาตัดคำตามกฎการประสมคำ 2 รูปแบบ ได้แก่ (1) คำที่ขึ้นต้นด้วยสระ (แ ไ โ ไ) และ (2) คำที่ขึ้นต้นด้วยพยัญชนะตัวควบกล้ำ (ร ล ว ทร) และพยัญชนะต้นที่มีตัวตาม (หม หน หล หย หน อย) รายละเอียดกฎแสดงในตารางที่ 1 และ 2 โดยกำหนดรหัสดังนี้ C แทนพยัญชนะต้น/ตัวควบกล้ำ/พยัญชนะต้นที่มีตัวตาม Vf แทน สระหน้า Vb แทนสระหลัง Vu แทน สระบน Vl แทนสระล่าง S แทน ตัวสะกด T แทนวรรณยุกต์/ไม้ไต่คู้ G แทนการันต์ โดยกฎโครงสร้างคำที่มีตัวการันต์ในตำรายาโบราณอีสานมีค่อนข้างน้อย เนื่องจากผู้ปริวรรตสะกดคำตามต้นฉบับ เช่น จันแดง จันขาว สนเท่ เป็นต้น ในงานวิจัยนี้จึงมีกฎสำหรับตัดคำที่มีตัวการันต์เพียงข้อเดียว และหากคำใดไม่ตรงกับกฎโครงสร้างข้อใดเลยจะคงคำนั้นไว้ดังเดิม ซึ่งมักเป็นคำหลายพยางค์ที่ยืมมาจากภาษาอื่น เมื่อแยกเป็นคำมักเป็นคำไม่ได้ความหมาย เช่น “ราชสีห์” “นกอินทรี” “สิทธิการิยะ” เป็นต้น หากต้องการจะตัดคำจะใช้วิธีตัดคำด้วยมือ แทน ผลจากการตัดคำทั้งในรอบที่ 1 และรอบที่ 2 จะตรวจสอบความถูกต้องโดยผู้เชี่ยวชาญอีกครั้ง หากยังตัดคำไม่ถูกต้อง หรือได้คำที่ไม่มีมีความหมาย จะแก้ไขด้วย

รูปที่ 3 ขั้นตอนและตัวอย่างการแปลงคำอ่านกลับเป็นคำปริวรรต

2.2.3 การทำคำปวิรรตของอักษรไทยน้อยด้วยวิธีแปลงคำอ่านภาษาไทยถิ่นอีสานกลับไปเป็นคำปวิรรตเพื่อใช้เป็นคำศัพท์หลักในคลังคำ กำหนดกฎดังนี้ (1) ตัดเครื่องหมายวรรณยุกต์ ตัวการ์นต์ และพยัญชนะควบกล้ำออกไปทั้งหมด เนื่องจากภาษาอีสานไม่มีตัวควบกล้ำ (2) คำอ่านที่ใช้พยัญชนะต้นหรือตัวสะกดที่ไม่นิยมใช้ในอักษรไทยน้อย (ข ฉ ญ ฎ ฏ ฐ ฑ ฒ ถ ภ ร ศ ษ ฟ ทร) ให้ใช้พยัญชนะต้นและตัวสะกดตรงตามเสียง ได้แก่ พยัญชนะต้น “ฆ” เปลี่ยนเป็น “ช”, “ฉ” และ “ศ” เปลี่ยนเป็น “ส”, “ทร” เปลี่ยนเป็น “ท”, “ช” เปลี่ยนเป็น “ซ” และ “ร” เปลี่ยนเป็น “ฮ” ส่วนพยัญชนะตัวสะกดแม้กกเปลี่ยนเป็น “ก” แมกกตเปลี่ยนเป็น “ด” แมกกบเปลี่ยนเป็น “บ” แมกนเปลี่ยนเป็น “น” และ (3) สระโอะไม่ผันเปลี่ยนเป็นสระโอะไม่มลาย (อักษรไทยน้อยไม่มีรูปสระโอะไม่ม้วน) สระอุ เปลี่ยนเป็น สระอู (อักษรไทยน้อยไม่มีรูปสระอู) จากนั้นนำคำปวิรรตที่ได้มาตรวจเทียบกับตารางต้นฉบับในเอกสารโบราณ หากไม่ตรงกันจะทำการแก้ไขให้ตรงกัน หากสงสัยหรือไม่แน่ใจจะสอบถามผู้เชี่ยวชาญ ตัวอย่างดังรูปที่ 3

การประเมินประสิทธิภาพของคลังศัพท์ แบ่งเป็น 2 ด้าน ดังนี้

3.2 ความเหมาะสมในการออกแบบและพัฒนาคลังศัพท์ ประเมินโดยนักภาษาโบราณ จำนวน 12 คน จากสำนักหอสมุดแห่งชาติและสถาบันวิจัยศิลปะและวัฒนธรรมอีสาน มหาวิทยาลัยมหาสารคาม และผู้ใช้ประโยชน์จากตำรายาโบราณอีสาน ได้แก่ นักศึกษาระดับปริญญาตรี ชั้นปีที่ 4 หลักสูตรแพทย์แผนไทยประยุกต์ คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม ที่ผ่านการเรียนรายวิชาตำรายาในเอกสารโบราณมาแล้ว จำนวน 27 คน ใช้แบบประเมินความเหมาะสมในการออกแบบ

$$\text{ร้อยละความถูกต้องของการแปลงคำปริวรรตเป็นคำอ่าน} = \frac{\text{คำอ่านที่แปลงได้ถูกต้อง}}{\text{จำนวนคำทั้งหมด}} \times 100 \quad (1)$$

$$\text{ร้อยละความผิดพลาดประเภทคำไม่รู้จัก} = \frac{\text{คำอ่านประเภทคำไม่รู้จัก}}{\text{จำนวนคำทั้งหมด}} \times 100 \quad (2)$$

$$\text{ร้อยละความผิดพลาดประเภทแปลงคำผิด} = \frac{\text{คำอ่านประเภทแปลงคำผิด}}{\text{จำนวนคำทั้งหมด}} \times 100 \quad (3)$$

$$\text{ผลรวมความผิดพลาด} = \text{จำนวนความผิดพลาดประเภทคำไม่รู้จัก} + \text{จำนวนความผิดพลาดประเภทแปลงคำผิด} \quad (4)$$

$$\text{ร้อยละความผิดพลาดทั้งหมด} = \frac{\text{ผลรวมความผิดพลาด}}{\text{จำนวนคำทั้งหมด}} \times 100 \quad (5)$$

ตารางที่ 3 จำนวนคำศัพท์ในคลังศัพท์ คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน ที่พัฒนาขึ้น

ประเภทความสัมพันธ์	จำนวน (คำ)	ร้อยละ
1:1	2,858	61.53
M:1	1,677	36.10
1:M	64	1.38
M:N	46	0.99
รวม	4,645	100

และพัฒนาฐานข้อมูลเป็นเครื่องมือในการรวบรวมข้อมูลใช้มาตรวัดแบบ Rating Scale 5 ระดับ

3. ผลการวิจัย

3.1 ลักษณะคลังศัพท์ภาษาไทยถิ่นอีสาน

โครงสร้างและลักษณะการอธิบายคำศัพท์ในคลังศัพท์ภาษาไทยถิ่นอีสานที่มีใช้ในปัจจุบัน ส่วนมากคล้ายคลึงกัน ดังนี้ คำศัพท์หลักพิมพ์ด้วยแบบอักษรไทยปัจจุบัน มีเพียง 1 เล่มที่มีคำศัพท์หลักพิมพ์ด้วยแบบอักษรไทยน้อย คือ พจนานุกรม 3 ภาษา (ไทยน้อย-อังกฤษ-ไทย) (Chantao, 2015) จำแนกชนิดของคำเป็น 7 ชนิด ได้แก่ คำนาม คำสรรพนาม คำกริยา คำวิเศษณ์ คำบุพบท คำสันธาน และคำอุทาน มีการอธิบายความหมายของคำศัพท์ มีตัวอย่างการใช้คำ (เฉพาะบางคำ) ไม่มีคำตรงข้ามและคำคล้ายคลึง ไม่มีคำปริวรรต คำศัพท์ส่วนใหญ่เป็นหมวดวรรณกรรมทั่วไป คำศัพท์หมวดตำรายาโบราณอีสานมีจำนวนน้อย และกระจัดกระจายอยู่ในบางเล่ม ทำให้ไม่สามารถนำคลังศัพท์ภาษาไทยถิ่นอีสานที่มีอยู่ในปัจจุบันไปใช้ในการตัดคำและแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยเครื่องได้ โดยคลังศัพท์ต้องมีลักษณะเฉพาะที่สำคัญคือ มีคำศัพท์หลักที่สะกดคำแบบคำปริวรรต มีคำอ่านภาษาไทยถิ่นอีสาน และหมวดคำศัพท์เป็นหมวดตำรายาโบราณอีสาน

3.2 คลังศัพท์ คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน มีลักษณะ ดังนี้

3.2.1 เป็นคลังศัพท์หมวดตำรายาโบราณอีสาน มีคำศัพท์หลักเป็นคำปริวรรตของอักษรไทยน้อย จำนวน 4,645 คำ จำแนกเป็น 4 ประเภทตามความสัมพันธ์ระหว่างคำปริวรรตและคำอ่านภาษาไทยถิ่นอีสานที่พบในงานวิจัยนี้ได้แก่ คำปริวรรต 1 คำ มีคำอ่าน 1 คำ (1:1) คำปริวรรต 1 คำ มีคำอ่านได้หลายคำอ่าน (1:M) คำปริวรรตหลายคำมีคำอ่านคำเดียวกัน (M:1) และคำปริวรรตแต่ละคำมีคำอ่านได้หลายคำและคำอ่านแต่ละคำอ่านมีคำปริวรรตได้หลายคำ (M:N) ดังตารางที่ 3 โครงสร้างฐานข้อมูลคลังศัพท์ ประกอบด้วย 4 ตาราง ได้แก่ คำศัพท์ (Vocabulary) เอกสารอ้างอิง (References) คำ ป ริ ว ร ร ต (Transliteration) และ คำ ป ริ ว ร ร ต - คำ อ า น ภาษา ไทย ถิ่น อีสาน (TransHasVocab) ความสัมพันธ์ระหว่างตารางในฐานข้อมูลแสดงในรูปที่ 4

ฐานข้อมูลคลังศัพท์จะรองรับคำปริวรรตและคำอ่านที่มีความสัมพันธ์ทั้ง 4 รูปแบบ โดยคำอ่านภาษาไทยถิ่นอีสานในตารางคำศัพท์แต่ละคำจะมี Vocab_id คนละตัว เช่น คำปริวรรต “เฮือน” มีคำอ่านทั้งหมด 2 คำ ได้แก่ “เฮือน” และ “เฮื้อน” จะมี Vocab_id คือ 1325 และ 1328 ตามลำดับ โดยรายละเอียดคำอธิบายโครงสร้างฐานข้อมูล (Data dictionary) คีย์หลัก (Primary key: PK) และคีย์นอก (Foreign key: FK) ของแต่ละตาราง แสดงในตารางที่ 4

3.2.2 ชนิดของคำ (Part-Of-Speech: POS) งานวิจัยนี้ จำแนกชนิดของคำในคลังศัพท์ออกเป็น 36 ชนิด (ตารางที่ 5) โดยใช้เกณฑ์การแบ่งชนิดของคำตาม

ตารางที่ 4 คำอธิบายโครงสร้างฐานข้อมูลคลังศัพท์คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน

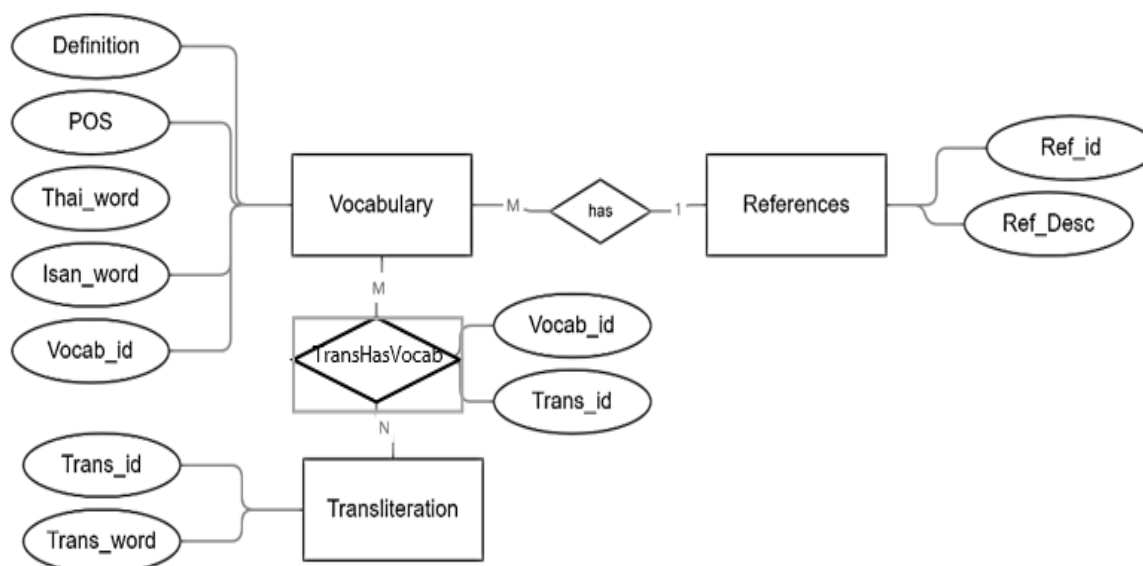
ชื่อแอททริบิวท์	ชนิดของข้อมูล	คำอธิบาย	ชื่อตาราง
Vocab_id (PK)	ตัวเลข (Integer)	รหัสคำศัพท์	คำศัพท์
Isan_word	ตัวอักษร (Text)	คำอ่านภาษาไทยถิ่นอีสาน	คำศัพท์
Thai_word	ตัวอักษร (Text)	คำอ่านภาษาไทยมาตรฐาน	คำศัพท์
POS	ตัวอักษร (Text)	ชนิดของคำ	คำศัพท์
Definition	ตัวอักษร (Text)	คำอธิบายคำศัพท์	คำศัพท์
Ref_id (FK)	ตัวเลข (Integer)	รหัสเอกสารอ้างอิง	คำศัพท์
Trans_id	ตัวเลข (Integer)	รหัสคำปริวรรต	คำปริวรรต
Trans_word	ตัวอักษร (Text)	คำปริวรรต	คำปริวรรต
Ref_id (PK)	ตัวเลข (Integer)	รหัสเอกสารอ้างอิง	เอกสารอ้างอิง
Ref_desc	ตัวอักษร (Text)	ชื่อเอกสารอ้างอิง	เอกสารอ้างอิง
Vocab_id (PK)	ตัวเลข (Integer)	รหัสคำศัพท์	คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน
Trans_id (PK)	ตัวเลข (Integer)	รหัสคำปริวรรต	คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน

ตารางที่ 5 คำอธิบายโครงสร้างฐานข้อมูลคลังศัพท์คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน

ลำดับ	POS	คำอธิบาย	ตัวอย่างคำศัพท์ภาษาไทยถิ่นอีสาน
1	NPRP	วิสามานยนาม (Proper Noun) หรือ คำนามประเภทชื่อเฉพาะอื่นที่ไม่ใช่ชื่อยา ชื่อโรค ชื่อสมุนไพร	ทิพยาธร, เทพพญาธร, ทิพพญาธร
2	NSMPT	คำนาม หรือ กลุ่มคำ (วลี) ประเภทชื่อโรค หรือ ชื่อเรียกอาการของโรค	สารบาท, สะดวง, ไข้มากไม้, ล่องแก้ว
3	NPMED	คำนามประเภทชื่อสมุนไพร (พืช, สัตว์, วัตถุ, กระจายยา) หรือ ส่วนประกอบของพืช หรือสัตว์ ที่นำมาใช้เป็นสมุนไพร	คันคาก (คางคก), อ้อยดำ, สากก้วยตีบ, ใบหมากเขียบ
4	NPTBE	คำนามประเภทชนิดของยา (Noun of Pharmacy Type)	น้ำ, ผง, ลูกกอน (ลูกกลอน), ฝุ่น (ผง)
5	NTME	คำนามประเภทเวลา หรือบอกเวลา (Nouns of Time)	เช้า (เช้า), แลง (ตอนเย็น), ตะวันชย (พระอาทิตย์ขึ้น), เพล (เที่ยง)
6	NDAY	คำนามประเภทชื่อวัน (Nouns of Day Name)	วันหนึ่ง (จันทร์), วันสอง (อังคาร), วันสาม (พุธ), ปะห์ส (พฤหัสบดี)
7	NDRT	คำนามประเภทชื่อทิศ (Noun of Direction)	ตะวันตก, ตะวันออก, เหนือ, ใต้
8	NORG	คำนามประเภทชื่อวัยวะ และชื่อสิ่งที่เกิดขึ้นภายในร่างกายมนุษย์ หรือสัตว์	แขน, ขา, มือ, เท้า, ฟันตึน, ฟันมือ, หน้าผาก, เลือด, เยี่ยว/เยี่ยว
9	NMTH	คำนามประเภทชื่อเดือนทางจันทรคติ	เดือน 1, อ้าย, ยี่, เจียง, เดือนสิบ
10	NCNM	คำนามประเภทจำนวนนับ หรือ จำนวนเชิงการนับ (Noun of Cardinal Number)	หนึ่ง, สอง, สาม, ..., 1, 2, 3, ...
11	NORD	คำนามบอกลำดับ	ชาว (ยี่), ชาวห้า (25), อ้อย (ร้อย)
12	PDMN	สามัญนาม (Common Noun) หรือ คำนามทั่วไปของ คน พืช สัตว์ และสิ่งของ	พ่อชย (ผู้ชาย), แม่ญิง (ผู้หญิง), เด็กน้อย, ม้า, หมากไม้, สาก, ใบ
13	PPRS	สรรพนามของคน (Personal Pronoun)	เพิ่น, ท่าน, เฮา, เจ้า, ไผ (ใคร)
14	PREL	ประพันธสรรพนาม หรือสรรพนามใช้แทนคำนาม หรือสรรพนามที่กล่าวมาแล้ว (Relative Pronoun)	มัน, อันนั้น
15	VERB	คำกริยาทั่วไป ที่ไม่ใช้กริยาในกลุ่มอาการป่วย วิธีเตรียมยา วิธีใช้ยา	ไป, ค้าง (วาง), ชุด, ให้, เอา, มี, ฐู (รู้), วาน (ขอร้อง), เห็น
16	VSTMP	คำกริยาประเภทอาการป่วย (Verb of Symptom)	ไค้ (บวม), พวน (มาน), ไอ, ออกฮ้อน (แสบร้อน), ฮ้อน (ร้อน), สิ้น
17	VMTD	คำกริยาประเภ่วิธีเตรียมยา (Verb of Method)	ฝน, บด, ตำ, ต้ม, แซ่ (แซ่), คั่ว
18	VDRCT	คำกริยาประเภ่วิธีใช้ยา (Verb of Direction)	กิน, อาบ, สม (รม), สูบ, โปะ (โปะ), พุด (พ่น), โสย (โรย)
19	XVRB	กริยาช่วย (Auxiliary Verb)	อย่า, เสีย, ไต้, สิ, จะ, ไว้
20	DCNM	คำบอกจำนวน (Determiner, Cardinal Number Expression)	ท่อกัน (เท่ากัน), อันแลอัน (อย่างละอัน), ซูอัน (ทุกอัน)
21	ADV N	คำกริยวิเศษณ์รูปปกติ (Adverb with Normal Form) ใช้ประกอบหรือขยายคำกริยาให้ได้ใจความชัดเจนและละเอียดมากขึ้น	ก่อน, ด้วย, แล้ว
22	ADVI	คำกริยวิเศษณ์รูปคำซ้ำ (Adverb with Iterative Form)	ก่อน2 (ก่อน ๆ), ทด2 (ทด ๆ)
23	ADJ N	คำคุณศัพท์รูปปกติ (Adjective with Normal Form)	แหม่ม (เกรียม), โหม้, ต่ำ, แดง, โหม้, ชุ่น, โหญ, น้อย
24	ADJI	คำคุณศัพท์รูปคำซ้ำ (Adjective with Iterative Form)	แหม่ม2 (แหม่ม ๆ), แดง2 (แดง ๆ), ชาว2 (ชาว ๆ), ชุ่น2 (ชุ่น ๆ)
25	CUNT	ชื่อหน่วยนับ (Unit Classifier)	ก้อน, ผ่าน, ซีก, บ่วง (ช้อน), บวย
26	CLTV	ลักษณะนามแบบกลุ่ม (Collective Classifier)	คู่, ชุม (พวก), ฝูง, ทั้งหมด, ทั้งหมด, ทั้งหมด

ตารางที่ 5 (ต่อ)คำอธิบายโครงสร้างฐานข้อมูลคลังศัพท์คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน

ลำดับ	POS	คำอธิบาย	ตัวอย่างคำศัพท์ภาษาไทยถิ่นอีสาน
27	CMTR	ลักษณะนามมาตรวัด (Measurement Classifier)	ตำลึง, บาท, จอมก้อย (ปลายนิ้วก้อย), เบี้ย, สลึง
28	CFQC	คำบอกลักษณะความถี่ (Frequency Classifier)	ที, เทื่อ (ครั้ง), คาบ (ครั้ง/เวลา)
29	CCJT	คำสัณฐาน หรือ คำเชื่อม (Coordinating Conjunction)	ผิว่า, และ, แล้วจึง, เป็นดัง, เพราะ, ถ้า, คั้น (หาก), คั้นว่า (หากว่า)
30	PREP	คำบุพบท (Preposition)	ใต้, ล่าง, เหนือ (บน), ใน, ทางใน, ทางลม, ทางหลัง
31	EAFF	คำจบประโยค (Ending for Affirmative Sentence)	เออะ, ดีแล, ดูก่อน, เทอญ, แล
32	NEG	คำปฏิเสธ (Negator)	บ่ (ไม่)
33	NEGI	คำปฏิเสธรูปซ้ำ (Negator with Iterative Form)	บ่2 (ไม่ ๑)
34	PUNC	เครื่องหมายวรรคตอน หรือจบประโยคตัวรับยา(Punctuation)	ฯ
35	SPELL	คาถา	อัติ วิมัสสมิง, สังคำ พันเต ปาวา
36	UNK	คำที่ระบุหมวดคำไม่ได้ ไม่แน่ใจ แปลไม่ได้	ลิสาย ทากวม สบสี สะหับ
27	CMTR	ลักษณะนามมาตรวัด (Measurement Classifier)	ตำลึง, บาท, จอมก้อย (ปลายนิ้วก้อย), เบี้ย, สลึง
28	CFQC	คำบอกลักษณะความถี่ (Frequency Classifier)	ที, เทื่อ (ครั้ง), คาบ (ครั้ง/เวลา)
29	CCJT	คำสัณฐาน หรือ คำเชื่อม (Coordinating Conjunction)	ผิว่า, และ, แล้วจึง, เป็นดัง, เพราะ, ถ้า, คั้น (หาก), คั้นว่า (หากว่า)
30	PREP	คำบุพบท (Preposition)	ใต้, ล่าง, เหนือ (บน), ใน, ทางใน, ทางลม, ทางหลัง
31	EAFF	คำจบประโยค (Ending for Affirmative Sentence)	เออะ, ดีแล, ดูก่อน, เทอญ, แล
32	NEG	คำปฏิเสธ (Negator)	บ่ (ไม่)
33	NEGI	คำปฏิเสธรูปซ้ำ (Negator with Iterative Form)	บ่2 (ไม่ ๑)
34	PUNC	เครื่องหมายวรรคตอน หรือจบประโยคตัวรับยา(Punctuation)	ฯ
35	SPELL	คาถา	อัติ วิมัสสมิง, สังคำ พันเต ปาวา
36	UNK	คำที่ระบุหมวดคำไม่ได้ ไม่แน่ใจ แปลไม่ได้	ลิสาย ทากวม สบสี สะหับ



รูปที่ 4 แผนภาพเอนทิตีและความสัมพันธ์ระหว่างเอนทิตีในฐานข้อมูลคลังศัพท์ คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน

วากยสัมพันธ์ (Syntax) ในภาษาไทยมาตรฐาน (Thonglor, 2000) และการแบ่งชนิดของคำตามโครงสร้างการเขียนตำราภาษาโบราณอีสาน

3.3 เกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสาน

เนื่องจากภาษาไทยถิ่นอีสานในแต่ละจังหวัดจะมีลักษณะเฉพาะ ดังนั้น คำศัพท์คำเดียวกันในบางจังหวัดอาจมีเสียงพยัญชนะ สระ หรือโทนเสียงวรรณยุกต์ต่างกัน (Thonglor, 2000) ทำให้การสะกดคำอ่านภาษาไทยถิ่นอีสานจะต้องมีการกำหนดเกณฑ์การสะกดคำตามสำเนียง

พูดในแต่ละจังหวัดด้วย โดยงานวิจัยนี้กำหนดเกณฑ์การสะกดคำอ่านตามสำเนียงพูดจังหวัดมหาสารคามและกาฬสินธุ์ (ตามผู้ปริวรรตตำราฉบับที่นำมาศึกษา) และไม่ใช่พยัญชนะ ควบกล้ำ เพื่อให้สอดคล้องกับลักษณะเฉพาะของสำเนียงภาษาไทยถิ่นอีสานที่ไม่มีเสียงพยัญชนะควบกล้ำ รวมถึงการพยายามรักษารูปแบบการใช้คำตามแบบโบราณเอาไว้ เพื่อประโยชน์ต่อการศึกษาด้านภาษาศาสตร์ และเป็นการสงวนรักษาวรรณกรรมด้านภาษาของแต่ละท้องถิ่นซึ่งเป็นมรดกทางวัฒนธรรมที่จับต้อง

ตารางที่ 6 เกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสานและตัวอย่างการสะกดคำอ่าน

เกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสาน	คำปริวรรต	ตัวอย่างการสะกดคำอ่าน	
		ภาษาไทยถิ่นอีสาน	ภาษาไทยมาตรฐาน
1. คำที่พ้องเสียงพ้องความหมายกับคำภาษาไทยมาตรฐาน ใช้การสะกดคำตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554	กิน	กิน	กิน
	เป็น	เป็น	เป็น
	ผักหวาน	ผักหวาน	ผักหวาน
2. คำที่ พ้องเสียงพ้องความหมายกับคำอ่านภาษาไทยมาตรฐานที่มีตัวการันต์ ใช้การสะกดคำตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554	จันทขาว	จันทขาว	จันทขาว
	ซิ่นกอินทรี	ซิ่นกอินทรี	ซิ่นกอินทรี
	นมลาซาฮี	นมราชสีห์	นมราชสีห์
3. คำที่ พ้องเสียงพ้องความหมายกับคำภาษาไทยมาตรฐานที่เป็น คำควบกล้ำ ร ล ว ใช้การสะกดคำตามเสียงภาษาไทยถิ่นอีสาน (ไม่ใช่พยัญชนะควบกล้ำ)	แพ	แพ	แพ
	ลูกกอน	ลูกกอน	ลูกกลอน
	กวาง	กวาง	กวาง
4. คำที่มีหน่วยเสียงพยัญชนะ /ส/ ที่เป็นปฏิภาคกับเสียงพยัญชนะ /ร/ ในภาษาไทยมาตรฐาน ใช้พยัญชนะ “ฮ” สะกดคำตรงกับเสียงภาษาไทยถิ่นอีสาน	เฮือน	เฮือน, เฮือน	เรือน, เรือน
	ฮาง	ฮาง, ฮ้าง	วาง, ร้าง
	ฮอน	ฮอน, ฮ้อน	ร้อน, ร้อน
5. คำที่มี หน่วยเสียงพยัญชนะ /ซ/ ที่เป็นปฏิภาคกับเสียงพยัญชนะ /ช/ ในภาษาไทยมาตรฐาน ใช้พยัญชนะ “ซ” สะกดคำตรงกับเสียงภาษาไทยถิ่นอีสาน	ซาง	ซ้าง	ช้าง
	ซุ่ม	ซุ่ม	ซุ่ม
	เชือก	เชือก	เชือก
6. คำภาษาไทยถิ่นอีสานที่คำปริวรรตประสมด้วยพยัญชนะต้น /ซ/ /ส/ /ท/ และประสมด้วยสระที่มีเสียงคล้ายกัน (สระเสียงสั้น/ยาว) แต่มีความหมายเหมือนกัน สะกดคำโดยใช้พยัญชนะและสระตรงกับเสียงภาษาไทยถิ่นอีสานสมัยปัจจุบัน	เสียบ, เขียบ	เสียบ	เสียบ
	สีก, สิก, ซีก	สีก, ซีก	ฉีก, ซีก
	ชะหน้อย, สน้อย	ชะหน้อย, สน้อย	สน้อย
7. คำภาษาไทยถิ่นอีสานที่คำปริวรรตประสมคำด้วย พยัญชนะ สระ เสียงต่างกัน แต่มีความหมายเดียวกัน การสะกดคำตามเสียงภาษาไทยถิ่นอีสานสมัยปัจจุบัน	เหลียง, เหลื่อง เหลื่อง, เหลียง	เหลื่อง	เหลื่อง
	นมลาซาฮี, นมลาฮี, นมละชะฮี	นมราชสีห์	นมราชสีห์
8. คำที่เป็นคำเฉพาะในภาษาไทยถิ่นอีสาน ใช้เกณฑ์การสะกดคำตามเสียงภาษาไทยถิ่นอีสานสมัยปัจจุบัน	หมากเขือบา	หมากเขือบ้า	ลำโพงกลัก
	คันคาก	คันคาก	คางคก
	หอยจูบ	หอบจูบ	หอยขม
9. คำที่ถอดมาจากพยัญชนะต้น “ย หางยาว” อักษรไทยน้อย การสะกดคำอ่านใช้รูปพยัญชนะต้นเป็น “ออย”	อยาก	อยาก	อยาก
	ออยู่	ออยู่	ออยู่
	ออย่าง	ออย่าง	ออย่าง
	ออย่า	ออย่า	ออย่า
10. คำที่มีหน่วยเสียงพยัญชนะต้นเป็นเสียง ย นาสิก ที่พ้องเสียงพ้องความหมายกับคำภาษาไทยมาตรฐาน สะกดคำตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 แต่หากเป็นคำศัพท์เฉพาะในภาษาไทยถิ่นอีสานให้สะกดด้วย “ญ”	หยา	หญา	หญา
	แมย้ง	แมญ้ง	ผู้ญ้ง
	ยอน	ญอน	เพราะ
	หยม	หญม	ขยม

ไม่ได้ (Intangible Cultural Heritage) ของชาติ สำหรับคำศัพท์ที่พ้องเสียงพ้องความหมายกับคำภาษาไทยมาตรฐานจะใช้เกณฑ์การสะกดคำตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 เพื่อให้อ่านเข้าใจได้ง่าย โดยกำหนดเกณฑ์การสะกด คำอ่านภาษาไทยถิ่นอีสานได้ 10 ข้อ ดังตารางที่ 6

จากตารางที่ 6 จะมีคำอ่านภาษาไทยถิ่นอีสานบางคำที่เป็นคำเดียวกันแต่มีคำอ่านต่างกัน ซึ่งเกิดจากมีคำปริวรรตเขียนสะกดต่างกัน การเก็บข้อมูลในฐานข้อมูลจะเก็บปริวรรตและคำอ่านภาษาไทยถิ่นอีสานไว้ทุกคำเพื่อให้สอดคล้องกับการออกเสียงพยัญชนะ และสระ ตามเอกสารต้นฉบับ

3.4. ผลการประเมินประสิทธิภาพคลังศัพท์

ผลการประเมินประสิทธิภาพคลังศัพท์คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-ภาษาไทยมาตรฐาน ในด้านการแปลงคำปริวรรตของตำรับยาโบราณอีสานที่บันทึกด้วยอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสาน (ตารางที่ 7) มีระดับความถูกต้อง (Accuracy) ร้อยละ 71.02 ประเภทที่มีความถูกต้องมากที่สุด คือ แบบ 1:1 ร้อยละ 51.72 รองลงมา คือ แบบ M:1 ร้อยละ 19.30 ส่วนความสัมพันธ์ประเภท 1:M และ M:N มีความถูกต้อง ร้อยละ 0 โดยเป็นความผิดพลาด ร้อยละ 28.98 แบ่งเป็นประเภท “คำที่ไม่รู้จัก” ร้อยละ 28.88 สาเหตุเกิดจากไม่มีคำปริวรรตดังกล่าวในคลังศัพท์ หรือคำปริวรรตมีคำอ่านมากกว่า 1 คำ ไม่สามารถเจาะจงคำอ่านที่ถูกต้องได้ และ “แปลงคำผิด” ร้อยละ 0.10 สาเหตุเกิดจากคำอ่านที่ได้มีความหมายไม่ถูกต้องตรงตามบริบท

ตารางที่ 7 ผลการประเมินความถูกต้องในการแปลงคำปริวรรตของตำรายาโบราณอีสานเป็นคำอ่านภาษาไทยถิ่นอีสานโดยใช้คลังศัพท์ คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน

ประเภทคำ ปริวรรต	จำนวนคำปริวรรต		ถูกต้อง		ผิดพลาด					
	ทดสอบ				คำไม่รู้จัก		แปลงคำผิด		รวม	
	คำ	ร้อยละ	คำ	ร้อยละ	คำ	ร้อยละ	คำ	ร้อยละ	คำ	ร้อยละ
1:1	4,396	55.44	4,101	51.72	292	3.68	3	0.04	295	3.72
M:1	1,775	22.39	1,530	19.30	240	3.03	5	0.06	245	3.09
1:M	1,308	16.50	0	0.00	1,308	16.50	0	0.00	1,308	16.50
M:N	450	5.68	0	0.00	450	5.68	0	0.00	450	5.68
รวม	7,929	100.00	5,631	71.02	2,290	28.88	8	0.10	2,298	28.98

ตารางที่ 8 ผลการประเมินประสิทธิภาพคลังศัพท์ด้านความเหมาะสมในการออกแบบและพัฒนาคลังศัพท์

ข้อ	รายการประเมิน	$\bar{X}$	SD	ประสิทธิภาพ
1	โครงสร้างคลังศัพท์มีความเหมาะสมกับการนำไปใช้แปลงคำปริวรรตเป็นคำอ่านภาษาไทยถิ่นอีสาน และคำอ่านภาษาไทยมาตรฐาน	4.08	0.67	ดี
2	คำศัพท์ในคลังศัพท์ครอบคลุมคำศัพท์ในหมวดตำรายาโบราณอีสาน	3.83	0.58	ดี
3	จำนวนคำศัพท์มีเพียงพอสำหรับใช้ปริวรรตตำรายาโบราณอีสาน	3.55	0.82	ดี
4	ความเหมาะสมของเกณฑ์การสะกดคำอ่านภาษาไทยถิ่นอีสาน	3.91	0.72	ดี
	ค่าเฉลี่ยทุกรายการ	3.89	0.71	ดี

ผลการประเมินประสิทธิภาพด้านความเหมาะสมในการออกแบบและพัฒนาคลังศัพท์ (ตารางที่ 8) พบว่าค่าเฉลี่ยทุกรายการของกลุ่มตัวอย่างทั้งสองกลุ่มเป็นไปในทิศทางเดียวกัน โดยค่าเฉลี่ยทุกด้านอยู่ในระดับดี ($\bar{X} = 3.89$, $SD = 0.71$) โดยรายการที่มีค่าเฉลี่ยมากที่สุด คือ โครงสร้างฐานข้อมูลคลังศัพท์มีความเหมาะสมสำหรับการนำไปใช้แปลงคำปริวรรตเป็นคำอ่านภาษาไทยถิ่นอีสานและคำอ่านภาษาไทยมาตรฐานอยู่ในระดับดี ($\bar{X} = 4.08$, $SD = 0.67$) และรายการประเมินที่มีค่าเฉลี่ยน้อยที่สุด คือ จำนวนคำศัพท์มีเพียงพอสำหรับใช้ปริวรรตตำรายาโบราณอีสานอยู่ในระดับดี ($\bar{X} = 3.55$, $SD = 0.82$)

4. วิจัยและสรุปผลการวิจัย

สรุปผลและวิจารณ์ผลการวิจัยได้ 5 ประเด็น ดังนี้

4.1 ฐานข้อมูลคลังศัพท์ “คำปริวรรต-คำอ่านภาษาไทยถิ่นอีสาน-คำอ่านภาษาไทยมาตรฐาน” หมวดตำรายาโบราณอีสาน ใช้วิธีรวบรวมคำศัพท์จากตำรายาโบราณอีสานที่ปริวรรตมาจากตำรายาโบราณอีสานที่บันทึกในใบลานด้วยอักษรไทยน้อย และตัดคำด้วยมือ มีข้อดีคือ ในกรณีหมวดคำศัพท์เฉพาะด้านจะได้คำศัพท์ที่ตรงกับความต้องการมากกว่าการรวบรวมจากคลังศัพท์ที่มีการสร้างเอาไว้แล้ว การรวบรวมคำศัพท์ใช้เวลาน้อยกว่า เนื่องจากไม่ต้องเสียเวลาในการค้นหาและคัดเลือกคำศัพท์ที่ต้องการจากแหล่งข้อมูลหลายแหล่ง สำหรับการนำประโยคในตำรายาโบราณอีสานมาตัดคำด้วยมือพบว่ามีความผิดพลาดต่ำกว่าการตัดคำด้วยโปรแกรมตัดคำภาษาไทย (Thai Word Segmentation) เช่น PyThaiNLP TLexPlus

หรือ Lexto เนื่องจากโปรแกรมเหล่านี้ไม่ได้สร้างจากฐานความรู้ทางภาษาที่เป็นภาษาไทยถิ่นอีสานโบราณ โดยเฉพาะ อย่างไรก็ตาม ระดับความถูกต้องในการตัดคำด้วยมือจะขึ้นอยู่กับความเชี่ยวชาญด้านภาษาศาสตร์ภาษาไทยถิ่นอีสาน และความรู้ด้านตำรายาอีสานของผู้สร้างคลังศัพท์เป็นหลัก นอกจากนี้ การตัดคำ 2 รอบ โดยรอบแรกตัดคำแบบยาวที่สุด และรอบที่ 2 นำคำประสมมาตัดคำแบบสั้นที่สุด โดยคำที่ตัดได้จะต้องเป็นคำที่มีความหมายนั้น ช่วยให้สามารถเพิ่มจำนวนคำศัพท์ในคลังศัพท์ได้มากขึ้น ทั้งนี้ เนื่องจากคำศัพท์ในตำรายาโบราณอีสานมีคำประสมอยู่ค่อนข้างมาก จึงสามารถนำมาแยกเป็นคำโดดที่มีความหมายแตกต่างกันได้หลายคำ

4.2 การพัฒนาคลังศัพท์อิเล็กทรอนิกส์ที่มีคำศัพท์หลักเป็นคำปริวรรต ช่วยลดข้อจำกัดด้านการไม่มีแบบอักษรโบราณ (Ancient Font) ที่ใช้พิมพ์คำศัพท์หลัก ซึ่งนอกจากจะนำไปใช้แปลงคำปริวรรตของอักษรไทยน้อยได้แล้ว ยังพบว่า สามารถนำไปใช้แปลงคำปริวรรตของอักษรธรรมอีสานได้อีกด้วย เนื่องจากเป็นอักษรโบราณที่คนไทยถิ่นอีสานใช้จดบันทึกตำรายาเช่นเดียวกับอักษรไทยน้อยโดยไม่จำเป็นต้องปรับเปลี่ยนโครงสร้างฐานข้อมูลและคำศัพท์ในคลังศัพท์แต่อย่างใด และการทำคำปริวรรตของอักษรไทยน้อยโดยใช้วิธีแปลงคำอ่านกลับเป็นคำปริวรรตพบว่า ช่วยให้ขั้นตอนการเตรียมคำศัพท์ทำได้เร็วกว่าการทำคำปริวรรตด้วยวิธีถอดรูปอักษรไทยน้อยจากเอกสารต้นฉบับโดยตรง ซึ่งการนำเกณฑ์การปริวรรตอักษรไทยน้อยเป็นอักษรไทยปัจจุบันมาใช้เป็นเกณฑ์ใน

การแปลงคำอ่านกลับเป็นคำปวิรรต ทำให้ผู้ที่มีความเชี่ยวชาญด้านการอ่านอักษรไทยน้อยในระดับพื้นฐานสามารถทำคำปวิรรตได้เร็วขึ้น

4.3 ประสิทธิภาพของการแปลงคำปวิรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยเครื่อง เนื้อหาหมวดตำรายาโบราณอีสาน แม้ว่าจะขึ้นอยู่กับขนาดของคลังศัพท์เป็นสำคัญ อย่างไรก็ตาม เนื่องจากคำปวิรรตที่มีสัดส่วนมากที่สุด 2 อันดับแรก เป็นคำปวิรรตที่มีความสัมพันธ์กับคำอ่านภาษาไทยถิ่นอีสาน แบบ 1:1 และ M:1 (ร้อยละ 61.53, และ ร้อยละ 36.10) เป็นกลุ่มคำที่ใช้เพียงคลังศัพท์ในการทำคำอ่านได้โดยไม่ต้องอาศัยคำบริบทที่อยู่โดยรอบมาช่วยทำคำอ่าน ประกอบกับในตำรายาโบราณอีสานมักมีการใช้คำเดิมซ้ำกันค่อนข้างมาก ดังนั้น แม้ว่าขนาดคลังศัพท์จะมีคำศัพท์เพียง 4,645 คำ แต่ประสิทธิภาพความถูกต้อง (Accuracy) ที่ได้ยังคงอยู่ในระดับปานกลางค่อนข้างดี (ร้อยละ 71.02)

4.4 ด้านความผิดพลาดของคลังศัพท์ที่มีอยู่ร้อยละ 28.98 พบว่า ส่วนใหญ่เกิดขึ้นกับคำปวิรรตประเภท 1:M และ M:N ทั้งสิ้น (มีความถูกต้องร้อยละ 0) โดยจำแนกความผิดพลาดได้เป็น 2 ประเภท ได้แก่ (1) “คำไม่รู้จัก” ซึ่งเกิดจากการที่ไม่มีคำปวิรรตนั้นในคลังศัพท์ หรือ “เป็นคำปวิรรตที่มีคำอ่านมีมากกว่า 1 คำ” ทำให้เจาะจงคำอ่านไม่ได้ และ (2) ความผิดพลาดแบบ “แปลงคำผิด” ที่เกิดกับคำปวิรรตประเภท 1:M ที่มีคำอ่านในคลังศัพท์เพียง คำเดียว สาเหตุเกิดจากตำรายาที่รวบรวมมาสร้างคลังศัพท์มีคำศัพท์ไม่ครอบคลุมคำศัพท์ที่เป็นไปได้ทั้งหมด ความผิดพลาดประเภทแปลงคำผิดจะทำให้คำอ่านมีความหมายไม่สอดคล้องกับคำบริบทที่อยู่รอบคำดังกล่าว เช่น “หมู” (คำปวิรรต) หากมีคำอ่านภาษาไทยถิ่นอีสานในคลังศัพท์เพียง 1 คำ คือ “หมู” เมื่อแปลงคำปวิรรตเป็นคำอ่านก็จะได้คำอ่านเป็นคำว่า “หมู” ในทุกประโยค ในขณะที่ในบางบริบทจะต้องมีคำอ่านเป็น “หมู” (พวก กลุ่ม) ไม่ใช่ “หมู” เป็นต้น ทั้งนี้ การลดความผิดพลาดกรณีดังกล่าว ชูดข้อมูลสำหรับสร้างคลังศัพท์จะต้องมีปริมาณมากและครอบคลุมคำอ่านภาษาไทยถิ่นอีสานที่เป็นไปได้ทั้งหมดของคำปวิรรตแต่ละคำ

4.5 ความแตกต่างของฐานข้อมูลคลังศัพท์ที่พัฒนากับคลังคำศัพท์ภาษาอีสานโบราณอื่น ๆ เมื่อเปรียบเทียบกับงานวิจัยที่คล้ายคลึงกัน 2 เรื่อง ได้แก่ การพัฒนาพจนานุกรมอักษรธรรมอีสานของ นงคณัฐ ไพบูลย์ (Paiboon, 2009) และโสภณ พรหมโสตา (Phromsoda, 2010) ทั้ง 2 พจนานุกรมเป็นพจนานุกรมสำหรับใช้แปลเอกสารโบราณที่บันทึกด้วยอักษรธรรมอีสาน เนื้อหา

หมวดวรรณกรรม นิทานชาดก และพุทธทำนาย โครงสร้างการอธิบายคำศัพท์ของพจนานุกรมทั้ง 2 พจนานุกรมมีความคล้ายคลึงกันคือ ประกอบด้วยคำศัพท์หลักอักษรธรรมอีสาน คำอ่านแปลภาษาไทย และความหมาย และจัดเก็บคำศัพท์ในโครงสร้างข้อมูลแบบทรี (Tries) ใช้แปลเอกสารโบราณที่พิมพ์ด้วยอักษรธรรมอีสานเท่านั้น ไม่สามารถนำใช้กับเอกสารโบราณอักษรไทยน้อยได้ ในขณะที่แนวคิดการพัฒนาคำศัพท์ที่มีคำศัพท์หลักเป็นคำปวิรรตที่ได้นำเสนอในงานวิจัยนี้ สามารถนำไปใช้แปลเอกสารโบราณอักษรธรรมอีสานและอักษรไทยน้อยได้ เพียงเพิ่มขั้นตอนการถ่ายถอดอักษรธรรมอีสานหรืออักษรไทยน้อยเป็นคำปวิรรต แล้วจึงนำคำปวิรรตไปค้นหาคำอ่านในคลังศัพท์ ดังนั้น การพัฒนาคำศัพท์ลักษณะนี้จึงช่วยลดการสร้างคลังคำศัพท์เฉพาะของอักษรโบราณแต่ละอักษรได้ รวมถึงสามารถนำแนวคิดนี้ไปใช้พัฒนาคำศัพท์สำหรับแปลเอกสารที่บันทึกด้วยอักษรต่างกันแต่บันทึกเป็นภาษาเดียวกันได้

อย่างไรก็ดี แม้ว่าจะระดับความถูกต้องในการตัดคำและการทำคำอ่านด้วยคลังศัพท์จะขึ้นอยู่กับจำนวนคำศัพท์ในคลังศัพท์ กล่าวคือ ยิ่งมีคำศัพท์มากก็ยิ่งเพิ่มความถูกต้องให้สูงเพิ่มมากขึ้น แต่สำหรับกรณีที่คำปวิรรตประเภท 1:M และ M:N นั้น ยังไม่สามารถใช้คลังศัพท์ที่มีขนาดใหญ่เพียงพอได้ แต่ต้องอาศัยคำบริบทที่อยู่รอบข้างมาช่วยเจาะจงคำอ่านที่ถูกต้องด้วย ดังนั้น ภายใต้โครงการวิจัยเดียวกันนี้ ผู้วิจัยได้ทำการศึกษาและนำเสนอขั้นตอนวิธีแปลงคำปวิรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานโดยใช้คลังศัพท์นี้ร่วมกับออนโทโลยี เพื่อเพิ่มประสิทธิภาพความถูกต้องในการทำคำอ่านให้เพิ่มสูงขึ้น

กิตติกรรมประกาศ

ขอขอบคุณ สถาบันวิจัยศิลปะและวัฒนธรรมอีสาน มหาวิทยาลัยมหาสารคาม ที่ให้ความอนุเคราะห์ข้อมูลตำรายาโบราณอีสานมาใช้ในการวิจัยนี้

เอกสารอ้างอิง

- Azizan, A., Jamal, N. N. S. A., Abdullah, M. N., Mohamad, M., & Khairudin, N. (2019). Lexicon-Based Sentiment Analysis for Movie Review Tweets. In *Conference, 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 132-136).
- Bianco, G., Bruno, F., Tonazzini, A., Salerno, E., Savino, P., Zitová, B., Sroubek, F., & Console E. (2010). A Framework for Virtual Restoration of Ancient Documents by Combination of Multispectral and 3D Imaging. In *Proceedings of Eurographics Italian Chapter Conference 2010* (pp. 1-7).

- Chamchong, R., Fung, C., & Chun, C. (2011). *Character Segmentation from Ancient Palm Leaf Manuscript in Thailand*. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing* (pp. 140–145).
- Chantao, R. (2015). *Dictionary of 3 Languages (Thai Noi-Lao-Thai)*. Khon Kaen University. (in Thai)
- Chawsuan, E. (2015). The combinations of consonants or syllables in words found in literature written in the Thai Noi script. *Silpakorn Journal*, 45–59. (in Thai)
- Chueaphun, C., Klomsae, A., Marukat, S., & Chaijaruwanich, J. (2012). Lanna Dharma Printed Character Recognition using k-Nearest Neighbor and Conditional Random Fields. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2012)* (pp. 169–174).
- Fang, S., & Xu, L. (2021, May). Creation and Significance of Database of Dictionary of Cognate Words. In *Workshop on Chinese Lexical Semantics* (pp. 119–129). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-06547-7_9
- Hollaus, F., Gau, M., & Sablatnig, R. (2012). Multispectral Image Acquisition of Ancient Manuscripts. In *Proceedings of Progress in Cultural Heritage Preservation 4th International Conference* (pp.30–39).
- Inkeaw, P., Charoenkwan, P., Huang, H.L., Marukat, S., Ho, S.Y., & Chaijaruwanich, J. (2017). Recognition of handwritten Lanna Dhamma characters using a set of optimally designed moment features. *Document Analysis and Recognition (IJDAR)*, 20(4), 259–274.
- Intasorn, J., Gertphol, S., & Sammapun, U. (2021). *Thai Sentiment Lexicon Construction*. In *Conference, 2021 13th International Conference on Knowledge and Smart Technology (KST)* (pp. 123–128).
- Kaewplang, K. (2013). Thai noi script: Language relation of Thai-Lao cultural. *Journal of Culture*, 52(4), 64–70. (in Thai)
- Khamluehan, C. (1992). Special orthography of the Inventive Dhamma script and Thai Noi. *Journal of Srinakharinwirot University Maha Sarakham campus*, 11(1), 33–38. (in Thai)
- Kingkham, W. (2013). *Thai dialects*. Kasetsart University printing house. (in Thai)
- Lorattanachaiyong, I. (2017). Generating Thai sentiment lexicon from online reviews [Unpublished master's thesis]. Chulalongkorn University. (in Thai)
- Na Nagara, P. (1998). Guidelines for translating Lanna native characters. *Journal of Prasert Na Nagara*, 71–73. (in Thai)
- Natarajan, J., & Sreedevi, I. (2017). Enhancement of ancient manuscript images by log based binarization technique. *AEU – International Journal of Electronics and Communications*, 75, 15–22.
- National Library of Thailand. (2002). *Collection of Lao language lexicals in ancient documents*. Prachachon. (in Thai)
- Office of the Royal Society. (2011). *Royal Institute Dictionary 2011* (2nd ed.). Nanme.
- Office of the Royal Society. (n.d.). *Terminology of the Office of the Royal Society*. <https://coined-word.orst.go.th>.
- Paiboon, N. (2009). *Isan Dhamma alphabet dictionary* [Unpublished master's thesis]. Khon Kaen University. (in Thai)
- Paiboonwangcharoen, P. (1999). Transliteration of ancient documents. *Manusat Paritat: Journal of Humanities*, 21, 25–30. (in Thai)
- Phimworamethakun, B. & Phimworamethakun, N. (2002). *Isan language dictionary: Speak Isan*. Khlang Nana Witthaya. (in Thai)
- Phromsoda, S. (2010). *Translation of Isan script into Thai* [Unpublished master's thesis]. Khon Kaen University. (in Thai)
- Pinthong, P. (1989). *Encyclopedia of Isan – Thai – English languages*. Siritham Printing House.
- Polsri, J. (2021). *The Development of a Framework for Transliterating Thai Noi Characters into Isan Language Pronunciation*. [Unpublished doctoral's thesis]. Technology Suranaree University. (in Thai)
- Punnotok, T. (2006). *Ancient Thai script, Lai Sue Thai script, and the evolution of Thai script*. Chulalongkorn University Printing House. (in Thai)
- Raksuthi, S. (2011). *Isan-Central Thai Dictionary*. Phatthana Sueksa. (in Thai)
- Romulus, P., Maraden, Y., Purnamasari, P. D., & Ratna, A. A. P. (2015). An analysis of optical character recognition implementation for ancient Batak characters using K-nearest neighbors' principle. In *the proceedings of 2015 International Conference on Quality in Research (QiR)* (pp. 47–50). Lombok, Indonesia.
- Sattanakho, K. (1999). Special orthography in Isan-Lanna script with stone inscription main I. *Maha Sarakham Journal*, 17(2), 64–70. (in Thai)
- Somdet Phra Maha Wirawong (Tissamah Thera). (1998). *Isan-central Thai dictionary Somdet Phra Maha Wirawong's aspirations edition (Tissamah Thera)*. Amarin Printing and Publishing. (in Thai)
- Somjitsripanya, S. (1982). Thai Noi script, Kab script, Soi script. Memorial of the royal funeral service of Phraratchasinsophit (Tha Phutthasorn Mahathera), the Ecclesiastical Provincial Governor of Mahasarakham Province. (in Thai)