การขาดมาตรฐานในการถอดอักษรไทยเป็นอักษรโรมันในการเขียนชื่อบุคคลไทยอย่างเหมาะสมทำให้การค้นหาชื่อบุคคลเป็นเรื่องที่ท้าทาย การถอดชื่อของบุคคลอย่างถูกต้องจะเป็นส่วนสำคัญในการค้นหาเอกสารที่เป็นภาษาอังกฤษที่เกี่ยวข้องกับบุคคลนั้นจากชื่อของบุคคลที่สะกดด้วยตัวอักษรไทยเพียงอย่างเดียว แต่การถอดอักษรบนพื้นฐานจากการออกเสียงชื่อของบุคคลเหล่านั้นโดยตรงมักจะนำไปสู่ความผิดพลาดจากการสะกดชื่อด้วยอักษรโรมันคนละแบบกับที่เจ้าของใช้เนื่องจากการสะกดด้วยอักษรไทยกับอักษรโรมันไม่ได้สัมพันธ์กันแบบ 1 ต่อ 1 ทั้งยังมีความนิยมส่วนบุคคลเข้ามาเกี่ยวข้องอีกด้วย งานวิจัยนี้เสนอวิธีการถอดอักษรโดยพิจารณาความนิยมในการใช้เข้ามาเกี่ยวข้อง โดยการแบ่งชื่อบุคคลไทยเป็นสายลำดับของแกรมซึ่งเป็นหน่วยย่อยที่ลักษณะคล้ายพยางค์ที่มีการบังคับจากระบบการเขียนและการออกเสียงทั้งจากภาษาไทยและภาษาอังกฤษ รวบรวมนำมาสร้างเป็นพจนานุกรมแกรมสะสมจากชื่อบุคคลไทย 130,000 ชื่อ ใช้แบบจำลองทางสถิติเข้ามาช่วยในการฝึกฝนบนพื้นฐานของแกรม เมื่อเปรียบเทียบกับวิธีการที่ใช้เป็นฐานซึ่งให้ผลความถูกต้องของการถอดอักษร 18 % วิธีการนี้ให้ผลที่ดีกว่าโดยให้ความถูกต้องของการถอด 46% - 75 % ของชื่อบุคคลที่สะกดอักษรโรมันเมื่อจำนวนของตัวเลือกที่จะเป็นคำตอบมากขึ้นจาก 1 ถึง 15

227278

The lack of standards for Romanization of Thai proper names makes searching activity a challenging task. This is particularly important when searching for people-related documents based on orthographic representation of their names using either solely Thai or English alphabets which is Roman based directly on the names' pronunciations often fails to deliver exact English spellings due to the non-1-to-1 mapping from Thai to English spelling and personal preferences. This paper proposes a Romanization approach where popularity of usages is taken into consideration. Thai names are parsed into sequences of grams, units of syllable-sized or larger governed by pronunciation and spelling constraints in both Thai and English writing systems. A Gram lexicon is constructed from a corpus of more than 130,000 names. Statistical models are trained accordingly based on the Gram lexicon. The proposed method significantly outperformed the current Romanization approach. Approximately 46% to 75% of the correct English spellings are covered when the number of proposed hypotheses increases from 1 to 15.