

รหัสชาวเด็กซ์แบบสมบูรณ์เพื่อการวิเคราะห์ความเหมือนของคำภาษาไทย

Complete Soundex for Thai Words Similarity Analysis

เฉลิมพล ทัพชัย (Chalermpol Tapsai)* พยุง มีสัจ (Phayung Meesad)*
และ ชูชาติ หฤไชยะศักดิ์ (Choochart Haruechaiyasak)**

Received : July 18, 2018
Revised : September 21, 2018
Accepted : October 10, 2018

บทคัดย่อ

งานวิจัยนี้เป็นการออกแบบและพัฒนาชาวเด็กซ์แบบสมบูรณ์ (Complete Soundex) ซึ่งเป็นวิธีการใหม่ที่จะต้องพิจารณาส่วนประกอบทั้งหมดของคำตามอักขรวิธีในการสร้างกฎการแปลงรหัส และนำเทคนิคการวิเคราะห์ความเหมือน (Similarity Comparison) มาใช้แทนการเปรียบเทียบรหัสตรงกัน (Matching Comparison) ของชาวเด็กซ์แบบเดิมๆ กระบวนการแปลงรหัสชาวเด็กซ์แบบสมบูรณ์ดำเนินการโดยเครื่องจักรแปลงรหัส (Converter Machine) ที่สร้างขึ้นจากกฎต่างๆ ที่ถูกออกแบบขึ้นใหม่โดยแบ่งออกเป็นกฎพื้นฐาน และกฎเพิ่มเติมที่ความสอดคล้องกับอักขรวิธีและการออกเสียงของส่วนประกอบแต่ละส่วนของคำ โดยมีตารางค่ารหัสของแต่ละส่วนประกอบโดยเฉพาะ การทดสอบและวิเคราะห์ประสิทธิภาพ กระทำโดยการป้อนชื่อและคำที่พิมพ์ผิดในลักษณะต่างๆ 6 รูปแบบได้แก่ การพิมพ์ไม่ครบ พิมพ์เกิน พิมพ์ซ้ำ พิมพ์ผิด พิมพ์ตัวอักษรสลับที่และพิมพ์สะกดแตกต่างกัน เข้าสู่ระบบชาวเด็กซ์แบบสมบูรณ์เพื่อค้นหาคำเหมือนจากพจนานุกรม โดยเปรียบเทียบผลลัพธ์กับวิธีชาวเด็กซ์แบบเดิม 4 วิธี ผลการทดสอบพบว่าชาวเด็กซ์แบบสมบูรณ์เป็นระบบที่มีประสิทธิภาพสูงที่สุด สามารถค้นหาคำที่ออกเสียงคล้ายกันพร้อมทั้งคำนวณค่าความเหมือนของคำซึ่งไม่มีในชาวเด็กซ์วิธีอื่น และมีคำเหมือนหลายคำที่ชาวเด็กซ์แบบเดิมไม่สามารถค้นได้ นอกจากนี้ยังสามารถแก้ปัญหาความผิดพลาดส่วนใหญ่ที่เกิดจากการความซับซ้อนในการใช้ตัวอักษรควบและสระ ซึ่งเป็นปัญหาที่ยังไม่ได้รับการแก้ไขในชาวเด็กซ์แบบเดิมด้วย

คำสำคัญ: ชาวเด็กซ์ รหัสเสียงสมบูรณ์ ความเหมือน
อัตโนมัติจำกัดเชิงไม่กำหนด

Abstract

This research aims to design and develop a Complete Soundex (CS), which is a new algorithm of entirely phonetic components encoding system, together with using of the similarity comparison technique instead of the matching comparison in Traditional Soundex (TS). The CS encoding process was performed by a converter machine and was created from the newly redesigned rules which were divided into basic rules and additional rules. All rules and tables of code values were defined corresponding to the phonetic feature of each word's component. Testing and evaluation are performed by input 6 types of misspelling words and names, including Excess of alphabets, Missing of alphabets, Repetition of alphabets, Typo error, Misplacement of alphabets, and alternative spelling, into the CS system to search for the similar words from the dictionary. By comparing with 4 TSs, the results show that the CS is the most effective system which can retrieve more precisely similar words with similarity values, which is not provided in the TSs. Moreover, It also found that the CS can solve most of the problems caused by the complexity of consonant clustering and vowels, that has not been fixed in the TSs.

* ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

* Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.

** หน่วยวิจัยวิทยาการสื่อสารของมนุษย์และคอมพิวเตอร์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

** Human Computer Communication Research Unit, National Electronics and Computer Technology Center.

Keywords: Soundex, Completely phonetic code, Similarity, Nondeterministic Finite Automata.

1. บทนำ

ชาวดึกซ์ (Soundex) เป็นระบบแปลงคำหรือชื่อต่างๆ ให้เป็นรหัสเสียงอ่าน เพื่อใช้เป็นเครื่องมือแก้ปัญหาคำที่ออกเสียงเหมือนกันแต่สามารถเขียนสะกดคำแตกต่างกันได้หลายแบบ ซึ่งจนถึงปัจจุบันมีการนำชาวดึกซ์ไปประยุกต์ใช้ในงานด้านต่างๆ ในหลายลักษณะ ได้แก่ ระบบการสืบค้นรายชื่อบุคคลจากสำมะโนประชากรชาวอเมริกัน ขององค์การบริหารจดหมายเหตุและบันทึกแห่งชาติ (National Archives and Records Administration : NARA) [1], การสืบค้นรายชื่อบุคคลในระบบติดตามเร่งรัดภาษีบำรุงท้องที่ของสำนักงานนโยบายและแผนกรุงเทพมหานคร [2], การค้นหาเอกสารเพื่อการเรียนรู้โดยใช้เทคนิคการวิเคราะห์ความหมายภาษาธรรมชาติ และใช้รหัสชาวดึกซ์เพื่อแนะนำคำถูกต้องเมื่อพบคำที่พิมพ์สะกดผิด [3], การแปลงชื่อข้ามภาษาเช่น การแปลงชื่อระหว่างภาษาจีนและอังกฤษโดยใช้รหัสชาวดึกซ์ ร่วมกับเทคนิคการทำ Web-mining [4], การแก้ไขคำสะกดผิดภาษาไทยโดยใช้รหัสชาวดึกซ์ [5], การแก้ไขคำสะกดผิดภาษาจีนโดยใช้รหัสชาวดึกซ์อักษรพินอิน [6], และระบบ Thai Q-Cor ที่สามารถแก้ไขคำสืบค้นภาษาไทยที่พิมพ์สะกดผิด โดยใช้เทคนิค Word approximation และใช้รหัสชาวดึกซ์เพื่อแก้ไขความผิดพลาดจากการพิมพ์สะกดที่แตกต่างจากคำในพจนานุกรมแต่ออกเสียงเหมือนกัน [7] เป็นต้น อย่างไรก็ตามในช่วงเวลาที่ผ่านมา มีงานวิจัยจำนวนมากที่เกี่ยวข้องกับการปรับปรุงกฎเกณฑ์ และกระบวนการแปลงรหัสชาวดึกซ์ด้วยเทคนิคและวิธีการต่างๆ แต่รหัสเหล่านี้ยังคงถูกออกแบบโดยตัดส่วนสำคัญที่ส่งผลต่อการออกเสียงของคำได้แก่ สระ วรรณยุกต์ และตัวควบกล้ำออกไป ซึ่งเมื่อนำไปเปรียบเทียบกับคำค้นที่รหัสตรงกัน มักส่งผลให้ไม่สามารถค้นพบคำที่ต้องการ และมีคำที่รหัสตรงกันจำนวนมากที่ไม่สามารถบอกความแตกต่างได้ว่า คำใดมีความเหมือนกับคำที่ใช้สืบค้นมากกว่าคำอื่นๆ ด้วยเหตุนี้ผู้วิจัยจึงมีความสนใจที่จะพัฒนาชาวดึกซ์แบบใหม่ที่สามารถแก้ปัญหาดังกล่าว รวมทั้งปรับปรุงกฎเกณฑ์การแปลงรหัส เพื่อแก้ปัญหาความผิดพลาดต่างๆ ที่เกิดขึ้นในวิธีการชาวดึกซ์แบบเดิม ซึ่งนอกจากจะเป็นการเพิ่ม

ประสิทธิภาพในการค้นคืนชื่อและคำต่างๆ ให้มีความถูกต้องมากขึ้นแล้ว ยังสามารถนำไปประยุกต์ใช้ในงานด้านอื่นๆ ที่เกี่ยวข้อง เช่น การประมวลผลภาษาธรรมชาติและ การแก้ไขคำที่พิมพ์สะกดผิดให้มีความถูกต้องมากยิ่งขึ้น

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ความเป็นมาของชาวดึกซ์

ชาวดึกซ์เกิดขึ้นครั้งแรกในปี ค.ศ. 1918 (พ.ศ. 2461) โดยชาวอเมริกันชื่อ Russell และ Odell ซึ่งได้จดสิทธิบัตร US Patents 1261167 (1918) และ 1435663 (1922) จากนั้นในอีก 2 ปีต่อมา จึงได้รับการพัฒนาให้เป็นรหัสชาวดึกซ์อเมริกัน (American Soundex) ที่ใช้สำหรับการวิเคราะห์ข้อมูลสำมะโนประชากรสหรัฐอเมริกา องค์การ NARA ได้จัดทำมาตรฐานการแปลงรหัส ชาวดึกซ์ สำหรับการใช้งานอย่างเป็นทางการจนถึงปัจจุบัน ซึ่งในช่วงเวลาต่อมา ชาวดึกซ์ได้ถูกพัฒนาและนำไปใช้งานจนเป็นที่รู้จักอย่างแพร่หลาย โดยถูกนำเสนอในบทความหลายฉบับของวารสาร Communications and Journal of the Association for Computing Machinery และมีการอธิบายโครงสร้างและการทำงานของรหัสชาวดึกซ์ ไว้ในหนังสือ ศิลปะการเขียนโปรแกรมคอมพิวเตอร์ ของโดนัลด์ นูธ ในปี ค.ศ. 1973

2.2 รหัสชาวดึกซ์ของ NARA

เว็บไซต์ของ NARA อธิบายเกี่ยวกับรหัสชาวดึกซ์ว่าเป็นรหัสที่สร้างขึ้นตามเสียงอ่านนามสกุลของชาวอเมริกัน ซึ่งถูกพัฒนาขึ้นเพื่อให้สามารถสืบค้นข้อมูลจากนามสกุลที่อ่านออกเสียงเหมือนกันได้ แม้จะสะกดแตกต่างกัน ซึ่งรหัสชาวดึกซ์นี้มีความยาวคงที่ 4 หลัก หลักแรกเป็นตัวอักษรที่ได้จากตัวอักษรตัวแรกของนามสกุล ส่วนรหัสอีก 3 หลัก ถัดมาเป็นตัวเลขที่ได้จากการเทียบตัวอักษรที่เหลือ กับกลุ่มตัวอักษรที่กำหนดไว้ใน Soundex Coding Guide [1] โดยในการแปลงรหัสนั้นหากมีตัวอักษรสองตัวที่อยู่ติดกันแปลงรหัสแล้วได้ตัวเลขเดียวกัน ให้ใช้ตัวเลขรหัสนั้นเพียงตัวเดียวสำหรับตัวอักษรที่ออกเสียงเป็นสระได้แก่ A, E, I, O, U, H และ W ให้ถือเป็นตัวค้น ไม่ต้องนำมาแปลงรหัส หากตัวค้นเป็น H หรือ W ให้ทั้งตัวเลขรหัสที่อยู่ทางขวา กรณีที่เป็นนามสกุลที่มีคำนำหน้า ได้แก่ คำว่า Van, Con, De, Di, La, หรือ Le สามารถแปลงรหัสได้ทั้งแบบที่มีและไม่มีคำนำหน้า ตัวอย่างเช่น

Washington แปลงเป็นรหัสชาวดั๊กซ์ W-252

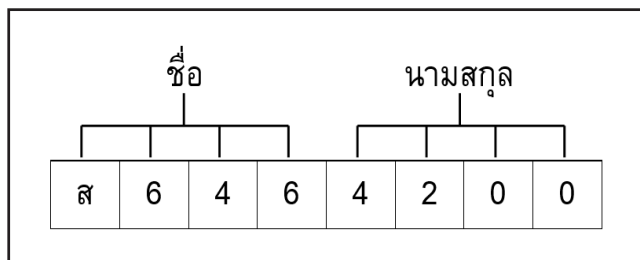
Gutierrez แปลงเป็นรหัสชาวดั๊กซ์ G-362

VanDeusen แปลงเป็นรหัสชาวดั๊กซ์ V-532, D-250

สำหรับภาษาไทยนั้น มีการนำเสนอวิธีการแปลงรหัสชาวดั๊กซ์ซึ่งเป็นที่รู้จักกันอย่างกว้างขวางจนถึงปัจจุบัน 4 วิธีได้แก่ วิธีของบริษัทดาต้าแมท [2] วิธีของวิจิต หล่อจิระชุมกุล [8] วิธีของวรรณิ อุดมพานิชย์ [9] และ วิธีของเทพพิทักษ์ การุญบุญญานันท์ และคณะ [5]

2.3 รหัสชาวดั๊กซ์ของบริษัทดาต้าแมท จำกัด

รหัสชาวดั๊กซ์ที่พัฒนาโดยบริษัทดาต้าแมท จำกัด เป็นรหัสที่สร้างขึ้นจากชื่อและนามสกุลของรายชื่อบุคคลในระบบติดตามเร่งรัดภาษีบำรุงท้องที่ของสำนักงานโยมบายและแผนกรุงเทพมหานคร เป็นรหัสที่มีขนาด 8 หลัก โดย 4 หลักแรกสร้างจากชื่อและอีก 4 หลักหลังสร้างขึ้นจากนามสกุลดังตัวอย่างภาพที่ 1



ภาพที่ 1 รหัสชาวดั๊กซ์ ของชื่อ สุวิทย์ ดีใจ

ในการแปลงรหัส ทั้งสระ วรรณยุกต์ และเครื่องหมายกำกับเสียงต่างๆจะถูกตัดทิ้งทั้งหมด แล้วใช้อักษรตัวแรกของชื่อเป็นรหัสหลักแรก ส่วนรหัสที่เหลืออีก 7 หลักเป็นตัวเลขที่ได้จากการแปลงรหัสตามตัวอักษรที่จัดไว้เป็นกลุ่มจำนวน 7 กลุ่ม โดยหากตัวอักษรในชื่อหรือนามสกุลหมดก่อนที่ตัวเลขในรหัสจะเต็ม ให้เติมศูนย์เข้าไปในรหัสส่วนที่เหลือ แต่หากได้ตัวเลขในรหัสครบแล้วให้ตัดตัวอักษรส่วนที่เกินทิ้งทั้งหมด

จะเห็นได้ว่าการแปลงรหัสชาวดั๊กซ์โดยวิธีนี้ ยังคงมีข้อด้อยที่ต้องพัฒนาปรับปรุงค่อนข้างมาก เนื่องจากเป็นการแปลงพยัญชนะทั้งหมดด้วยคำรหัสเพียงชุดเดียวโดยไม่แยกเป็นพยัญชนะต้นคำและตัวสะกดตามหลักภาษา ทั้งที่มีการออกเสียงที่ต่างกัน นอกจากนี้ยังไม่มีกรึงพยัญชนะที่มีการันต์ (์) กำกับอยู่ออกไป แม้จะเป็นพยัญชนะที่ไม่มีกรึงออกเสียงใดๆ จึงทำให้รหัสที่ได้ไม่ตรงกับการออกเสียงของชื่อที่นำมาเข้ารหัส

2.4 รหัสชาวดั๊กซ์ของวิจิต หล่อจิระชุมกุล

การแปลงรหัสชาวดั๊กซ์ด้วยวิธีนี้ มีการปรับปรุงกฎการแปลงรหัสเพื่อให้ได้รหัสที่มีความถูกต้องมากขึ้น เช่น การทิ้งตัวอักษรที่อยู่ใต้การันต์ และนำสระบางตัวเช่น ี ื ุ มาพิจารณาพร้อมกับพยัญชนะที่อยู่ในตำแหน่งเดียวกัน เพื่อตัดพยัญชนะบางตัวที่ไม่ต้องการนำมาแปลงรหัสทิ้งไป โดยได้ออกแบบให้รหัสมีความยาว 5 หลัก หลักแรกได้จากการแปลงพยัญชนะตัวแรกของชื่อให้เป็นตัวอักษร ซึ่งแบ่งออกเป็น 19 กลุ่มตามเสียงของพยัญชนะต้นคำ ส่วนหลักที่เหลือได้จากการแปลงอักษรตัวต่อๆ ไปในชื่อให้เป็นรหัสเลขฐาน 16 ตามเสียงของตัวสะกดและสระบางตัวที่ถูกจัดเป็นกลุ่มรวม 15 กลุ่ม ตัวอย่างรหัสชาวดั๊กซ์ของชื่อ-นามสกุลจริงเปรียบเทียบกับรหัสชาวดั๊กซ์ของคำอ่านที่สะกดให้ออกเสียงอย่างเดียวกันแสดงให้เห็นดังตารางที่ 1

ตารางที่ 1 ตัวอย่างรหัสชาวดั๊กซ์ที่สร้างขึ้นโดยวิธีของวิจิต หล่อจิระชุมกุล

ชื่อ-นามสกุลจริง	รหัสชาวดั๊กซ์	คำอ่านที่สะกดออกเสียงตามชื่อ	รหัสชาวดั๊กซ์
คณิศร์ เลิศศิริวิบูลย์	กบ430	คะเนต เลิศศิริวิบูล	กบ430
เบญจวรรณ ปราโมชกุล	บบ437	เบ้นจะวัน ปราโมตกุล	บบ437
สุพัตรา ศรีทอง	ชเ534	สุพัตตา สีทอง	ชเ530*
ธรรมชัย บุญนาย	ท4637	ท่าชัย บุนนาย	ท6370*

*จากตารางจะพบว่าการแปลงรหัสชาวดั๊กซ์โดยวิธีนี้ ยังคงมีปัญหาความผิดพลาดของรหัสชาวดั๊กซ์ที่ได้จากชื่อ-นามสกุลจริง ที่มีการสะกดในบางลักษณะ เช่นการใช้ “รร” ซึ่งออกเสียง “อัม” และการใช้ตัวสะกดมาตราแม่กดแบบตัวควบเช่น ชร ตร ตี รด รท เป็นต้น

2.5 รหัสชาวดั๊กซ์ของวรรณิ อุดมพานิชย์

วรรณิ อุดมพานิชย์ได้ออกแบบกฎการแปลงรหัสชาวดั๊กซ์สำหรับชื่อคนไทย ที่ผู้ใช้สามารถกำหนดความยาวของรหัสได้ตามความเหมาะสมในการใช้งาน โดยนำชื่อและนามสกุลมาต่อกัน และทำการแปลงเป็นรหัสต่อเนื่องกันไป นอกจากนี้ยังได้ปรับปรุงตารางแปลงรหัสพยัญชนะตัวแรกโดยแยกตัวอักษร ก ออกมาจากกลุ่มตัวอักษร ข, ช, ค, ฉ ทำให้จำนวนชุดของตัวอักษรที่เป็นพยัญชนะตัวแรกของ

คำเพิ่มขึ้นเป็น 20 กลุ่ม และตัดสระต่างๆ ออกจากชุดอักษรตัวตาม คงเหลือไว้เฉพาะสระ ำ พร้อมทั้งจัดกลุ่มใหม่ให้เหลือเพียง 10 กลุ่ม ตัวอย่างรหัสชาวดึกซ์ของชื่อ-นามสกุลจริง และคำอ่านที่สะกดให้อ่านออกเสียงตามชื่อแสดงดังตารางที่ 2

ตารางที่ 2 ตัวอย่างรหัสชาวดึกซ์ที่สร้างขึ้นโดยวิธีของวรรณี อุดมพาณิชย์

ชื่อ-นามสกุลจริง	รหัสชาวดึกซ์	คำอ่านที่สะกดออกเสียงตามชื่อ	รหัสชาวดึกซ์
คณิศร์ เลิศศิริวิบูลย์	ค34946900	คะเนด เลิศหิรันวิบูล	ค34946900
สุวัฒน์ สิงห์ทอง	ส08428600	สุวัด สิงทอง	ส08428600*
เบญจวรรณ ปราโมชกุล	บ37099300	เบ็นจะวัน ปราโมตกุล	บ37035900*

จากตารางจะพบว่าการแปลงรหัสชาวดึกซ์โดยวิธีนี้ ยังคงมีปัญหาความผิดพลาดค่อนข้างมาก ซึ่งเกิดจากสาเหตุต่างๆ เช่น การจัดกลุ่มอักษรตัวตามที่ไม่สอดคล้องกับการทำหน้าที่เป็นตัวสะกดในการออกเสียง การใช้ “รร” การที่มีตัวสะกดบางตัวออกเสียงเป็นพยัญชนะต้นของพยางค์ถัดไป การใช้ตัวอักษรควบ และการออกเสียง “ไอ” ที่สามารถเขียนสะกดได้หลายแบบ เป็นต้น

2.6 รหัสชาวดึกซ์ของเทพพิทักษ์ การบุญบุญนันท์ และคณะ

รหัสชาวดึกซ์ที่สร้างขึ้นโดยวิธีของเทพพิทักษ์ การบุญบุญนันท์และคณะ มีจุดมุ่งหมายเพื่อใช้ในการแก้ปัญหาการพิมพ์สะกดคำผิด เป็นรหัสที่ครอบคลุมคำอ่านที่เป็นไปได้ทุกๆ แบบ ตัวอย่างเช่นคำว่า กมล สามารถอ่านได้ทั้ง กะ-มน, กะ-มะ-ละ, กะ-มน-ละ, กอ-มอ-ลอ ฯลฯ โดยกฎแต่ละข้อถูกเขียนขึ้นตามหลักอักขรวิธีของการสร้างพยางค์ภาษาไทย ในรูปแบบของ Regular expressions with output มีการกำจัดส่วนที่เป็นตัวอักษรควบ (Clustering), วรรณยุกต์ (Tone) และความสั้นยาวของเสียงสระ (Vowel Length) ออกก่อนที่จะทำการแปลงรหัส ในการแปลงรหัส จะนำกฎทั้งหมดมาเขียน Nondeterministic Finite Automaton with Output เพื่อใช้สำหรับออกแบบการทำงานของเครื่องจักรแปลงรหัส (Converter Machine) โดยแบ่งตัวอักษรที่ต้องแปลงรหัสออก

เป็นทั้งหมด 3 ชุด ชุดแรกแบ่งตามเสียงพยัญชนะต้นคำ (Initial Consonant Sound) ซึ่งแบ่งออกเป็น 20 กลุ่ม ชุดที่สองแบ่งตามเสียงตัวสะกด (Final Consonant Sound) มี 9 กลุ่ม และชุดที่สามแบ่งตามเสียงสระ (Vowel Sound) ซึ่งแบ่งออกเป็น 12 กลุ่ม ตัวอย่างของรหัสที่สร้างขึ้นจากคำศัพท์ต่างๆ ด้วยวิธีการนี้แสดงในตารางที่ 3

ตารางที่ 3 ตัวอย่างรหัสชาวดึกซ์ที่สร้างขึ้นโดยวิธีของเทพพิทักษ์ การบุญบุญนันท์

คำศัพท์	รหัส ชาวดึกซ์
ธรรมชาติ	ทามมฺาฮ, ทานมฺาฮ, ทฺนลอมฺมาฮ, ทฺนลคฺสมฺาฮ, ทฺนลลสมฺาฮ, ทฺชลลคฺลอมฺมาฮ, ทฺชลลคฺลอมฺมาฮ, ทฺชลลคฺลอมฺมาฮ, ทฺชลลคฺลอมฺมาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ, ทฺชลลคฺลคฺสมฺาฮ
กระยิก	กาสยิก, กาสยิกกาส, กคฺลลคฺยิกกาส, กคฺลลคฺยิกกาส, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค, กคฺลลคฺยิกกค

จากตัวอย่างจะเห็นว่า การแปลงรหัสด้วยวิธีนี้มีการสร้างรหัสออกมาจำนวนมาก ซึ่งส่งผลให้การเปรียบเทียบเพื่อหาคำที่สะกดถูกต้องใช้เวลาเพิ่มขึ้นด้วย นอกจากนี้ การที่ไม่นำตัวอักษรควบ เสียงวรรณยุกต์และความสั้นยาวของเสียงสระเข้ามาพิจารณาในการแปลงรหัสก็ส่งผลให้รหัสที่ถูกสร้างขึ้นขาดความสมบูรณ์และไม่สามารถแยกความแตกต่างระหว่างคำที่มีรหัสตรงกันได้ ตัวอย่างเช่น เมื่อนำคำที่สะกดผิด “ปราณ” ซึ่งมีรหัสชาวดึกซ์หลายรหัสเช่น ปาน, ปานยาฮ ฯลฯ ซึ่งเมื่อนำไปเปรียบเทียบกับรหัสชาวดึกซ์ของคำอื่นๆ จะพบคำที่มีรหัส ปาน เหมือนกันจำนวนมากได้แก่คำว่า ปาน ปาน ปาน ปราณ ปราณ ปัน ปัน ปัน ปัน ปัน ปัน ปัน เป็นต้น

2.7 เปรียบเทียบการแปลงรหัสชาวดึกซ์ภาษาไทยแบบเดิม

จากกฎเกณฑ์และวิธีการแปลงรหัสชาวดึกซ์ทั้งสี่วิธีที่ได้นำเสนอข้างต้น สามารถนำมาสรุปเปรียบเทียบการใช้ส่วนประกอบต่างๆ ของคำมาใช้ในการแปลงรหัสของแต่ละวิธีได้ดังตารางที่ 4

ตารางที่ 4 การใช้ส่วนประกอบต่างๆของคำในการแปลงรหัส
ชาวเด็กรหัสภาษาไทยแบบเดิม

วิธีแปลงรหัส	พยานชนะต้นคำ	พยานชนะตัวตาม	สระ	วรรณยุกต์, ตัวควบกล้ำ
ดาต้าแมท	แบ่งเป็น 7 กลุ่มตามฐานฐานอวัยวะออกเสียง	ใช้รหัสชุดเดียวกันกับพยานชนะต้นคำ	ไม่ใช้	ไม่ใช้
วิจิต	แบ่งเป็น 19 กลุ่มตามเสียงพยานชนะต้น	ใช้พยานชนะและสระบางตัวแบ่งเป็น 15 กลุ่ม	ใช้บางตัว	ไม่ใช้
วรรณี	แบ่งเป็น 20 กลุ่มตามเสียงพยานชนะต้น	แบ่งเป็น 9 กลุ่มตามตัวสะกด	เฉพาะสระ ำ	ไม่ใช้
เทพพิทักษ์	แบ่งเป็น 20 กลุ่ม	แบ่งเป็น 9 กลุ่มตามตัวสะกด	แบ่งเป็น 12 กลุ่มตามเสียงสระ	ไม่ใช้

3. วิธีดำเนินการวิจัย

จากการศึกษาวิธีการแปลงรหัสด้วยวิธีต่างๆ ดังที่กล่าวมาข้างต้น พบว่าทุกวิธีเป็นการเปรียบเทียบเพื่อหาค่ารหัสชาวด์เด็กซ์ที่มีค่าตรงกันเท่านั้น จึงทำให้ไม่สามารถแยกความแตกต่างได้ว่าคำใดมีความเหมือนมากน้อยกว่ากันได้ นอกจากนี้ในแต่ละวิธียังมีข้อผิดพลาดและข้อด้อยอีกจำนวนมากที่ต้องได้รับการปรับปรุงเพื่อเพิ่มประสิทธิภาพและความถูกต้องของผลลัพธ์ให้ดีขึ้น เช่น ความผิดพลาดจากการใช้ตัวอักษรควบ เสียวรรณยุกต์ การใช้ตัวอักษร และรร ซึ่งการแก้ปัญหาเหล่านี้ จะต้องทำการปรับปรุงกฎและจัดกลุ่มตัวอักษรสำหรับการแปลงรหัส ให้สอดคล้องกับหลักอักขรวิธีของส่วนประกอบแต่ละส่วนที่ทำให้เกิดเสียงในแต่ละพยางค์ของคำ ซึ่งผู้วิจัยได้ออกแบบและพัฒนาชาวด์เด็กซ์แบบสมบูรณ์ โดยมีการดำเนินงานแบ่งออกเป็น 3 ขั้นตอนคือ การวิเคราะห์ลักษณะและหน้าที่ของส่วนประกอบต่างๆ ของคำ การออกแบบรหัสชาวด์เด็กซ์แบบสมบูรณ์ และการทดสอบและประเมินประสิทธิภาพ

3.1 การวิเคราะห์ลักษณะและหน้าที่ของส่วนประกอบต่างๆ ของคำ

นอกจากหลักอักขรวิธีพื้นฐานที่เกี่ยวข้องกับการเขียนคำภาษาไทยแล้ว ยังมีหลักสำคัญอื่นๆ ที่ต้อง

พิจารณาเพิ่มเติมเกี่ยวกับลักษณะและหน้าที่ของส่วนประกอบต่างๆ ที่ทำให้เกิดเสียงในคำ ซึ่งผู้วิจัยได้ทำการศึกษาเนื้อหาในเรื่องดังกล่าวจากหนังสือหลักภาษาไทยของพระยาอุปทิศศิลปสาร [10] และ กำชัย ทองหล่อ [11] แล้วสรุปเป็นหลักสำคัญๆ แยกอธิบายเป็นแต่ละหัวข้อดังต่อไปนี้

ส่วนประกอบต่าง ๆ ของคำภาษาไทย

คำในภาษาไทยประกอบขึ้นด้วยตัวอักษรต่างๆ 3 ชนิด คือพยัญชนะ สระและวรรณยุกต์ โดยแต่ละคำอาจมีการออกเสียงเพียง 1 พยางค์หรือหลายพยางค์ก็ได้ และอาจเป็นคำมูล (คำที่มีความหมายในตัวเอง) หรือเป็นคำประสมที่เกิดจากการนำคำมูลมากกว่า 1 คำมารวมกัน

พยานชนะ

พยัญชนะในภาษาไทยมีทั้งหมด 44 ตัวทำหน้าที่เป็นพยัญชนะต้นของคำหรือเป็นตัวสะกด โดยพยัญชนะส่วนใหญ่มีการออกเสียงต่างกันเมื่อทำหน้าที่เป็นพยัญชนะต้นหรือเป็นตัวสะกด การใช้พยัญชนะในภาษาไทยมีทั้งลักษณะเป็นตัวอักษรเดี่ยวและเป็นตัวอักษรควบ และอาจมีเครื่องหมายวรรคตำแหน่งกับหากไม่ต้องการให้มีการออกเสียงพยัญชนะตัวดังกล่าว

สระและวรรณยุกต์

สระในภาษาไทยมีทั้งหมด 32 ตัวแบ่งออกเป็นสระเสียงสั้น 18 ตัวได้แก่ ะ, ิ, ี, ุ, ึ, แะ, ๅ, โะ, ๊, ๋, ์, ํ, ๎, ๏, ๐ และสระเสียงยาว 14 ตัวได้แก่ ำ, ั, ็, ู, ื, ๑, ๒, ๓, ๔, ๕, ๖, ๗, ๘ โดยเสียงของสระบางตัวมีเสียงตัวสะกดกำกับด้วย ได้แก่

สรระ ำ ประกอบด้วยเสียงสรระ ะ และตัวสะกด ม

สระ ใ และ ใ ประกอบด้วยเสียงสระ ะ และตัวสะกด ย

สระ ๔ ประกอบด้วยเสียงสระ ะ และตัวสะกด ว

นอกจากนี้ สาระบางตัวในภาษาไทย มีการแปลงรูปในกรณีต่างๆ ได้แก่

สระ ะ เมื่อมีตัวสะกดจะแปลงรูปเป็น ั ตัวอย่างเช่น คำว่า กัด แปลงรูปมาจาก ก-ะ-ด เป็นต้น

สระ เ เอ เมื่อมีตัวสะกดจะแปลงรูปเป็น เ ตัวอย่างเช่น
เป็น แปลงรูปมาจาก ป-เ-น เป็นต้น

สระ แ และ อ เมื่อมีตัวสะกดจะแปลงรูปเป็น แอ ตัวอย่างเช่น
แข็ง แปลงรูปมาจาก ข-แ และ-ง เป็นต้น

สระ อ แปลงรูปเป็น ใ มีใช้เพียงคำเดียวคือคำว่า ก็ แปลง

รูปมาจาก ก้อ

วรรณยุกต์

ในภาษาไทยใช้วรรณยุกต์เป็นเครื่องหมายกำกับเสียงซึ่งมีทั้งหมด 5 เสียง 4 รูป ได้แก่ เสียง สามัญ (ไม่มีรูป) เอก (ˊ) โท (ˊˊ) ตรี (ˋ) และจัตวา (ˋˋ)

อักษรควบ

คือพยัญชนะ 2 ตัวที่เขียนควบหรือล้ากันอยู่ในสระตัวเดียวกัน โดยมีตัวหนึ่งเป็น ร, ล, ห หรือ ว อักษรควบมี 2 ชนิดคือ อักษรควบแท้ และ อักษรควบไม่แท้

อักษรควบแท้ เป็นการควบระหว่างพยัญชนะตัวหนึ่งกับตัว ร, ล หรือ ว สามารถทำหน้าที่ได้สามกรณีคือ

1) เป็นพยัญชนะต้นคำ โดยออกเสียงพยัญชนะทั้งสองตัวควบกัน เช่นคำว่า ครู ไกร กลอง ขวาง เป็นต้น

2) เป็นตัวสะกดได้แก่ กร คร ตร ทร และ พร เช่นคำว่า จักร สมักร บัทร สมุทร กอปร ซึ่งในการออกเสียงเป็น ตัวสะกดนี้จะไม่ออกเสียงตัว ร แต่หากมีสระกำกับด้านหลัง จะต้องออกเสียงเป็นทั้งตัวสะกด และเป็นทั้งพยัญชนะต้นของพยางค์หลังด้วย เช่น บุตรี ออกเสียงเป็น บุด-ตรี, จักรี ออกเสียงเป็น จัก-กรี เป็นต้น

3) เป็นตัวการันต์ด้วยกัน เช่น พักตร์ ศาสตร์ อินทร์ อัมรินทร์ เป็นต้น

อักษรควบไม่แท้ เป็นการควบระหว่างพยัญชนะตัวหนึ่งกับตัวอักษร ร โดยตัวอักษร ร จะอยู่หน้าหรือหลังก็ได้ สามารถทำหน้าที่ได้ 4 กรณีคือ

1) ควบกับพยัญชนะต้นโดยไม่ออกเสียงตัว ร เช่นคำว่า จริง ศรีทธา เศร้า สร้อย เสรี เป็นต้น

2) ออกเสียงเปลี่ยนไปเป็นเสียงตัวอื่นได้แก่ ตัวอักษรควบทร ออกเสียงเป็น ซ เช่นคำว่า ทราย ทราบ เป็นต้น

3) ใช้เป็นตัวสะกด ได้แก่ ตัวอักษร ร และ ห ที่ควบกับพยัญชนะตัวอื่น โดยเมื่อเขียนควบกันแล้วจะไม่ออกเสียงตัวอักษร ร และ ห ยกเว้นกรณี ของตัวอักษรควบ รร ที่สามารถออกเสียงได้สองอย่างคือ กรณีที่ รร ตามด้วยตัวสะกด ให้ออกเสียงเป็นสระ อะ แต่หากไม่มีตัวสะกด ให้ออกเสียงเป็น อัน เช่นคำว่า พรรด ออกเสียงเป็น พัก, สรร ออกเสียงเป็น สัน, จรรยา ออกเสียงเป็น จัน-ยา, มารค ออกเสียงเป็น มาก เป็นต้น

4) เป็นตัวการันต์เช่นคำว่า ครรรจ์ ธรรม์ เป็นต้น

อักษรนำ

คือพยัญชนะ 2 ตัวที่เขียนร่วมอยู่ในสระตัวเดียวกัน แบ่งออกเป็น 2 ชนิดคือ

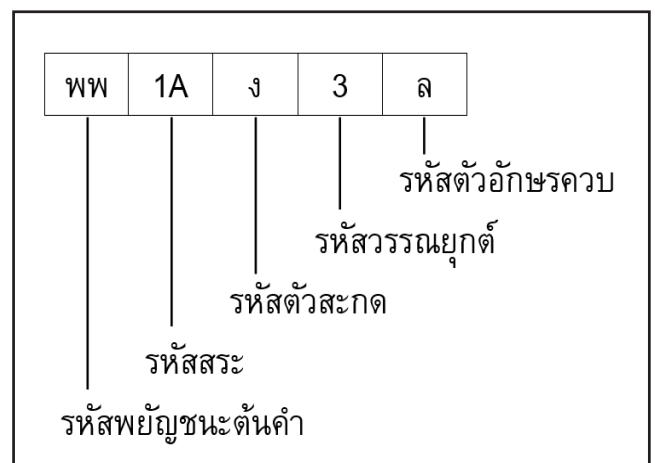
1) ชนิดที่ออกเสียงร่วมกันสนิท ได้แก่อักษรนำที่มีตัว ห และ อ นำหน้า เช่น หนู หมอ อย่า อยู่ อยาก เป็นต้น

2) ชนิดที่ออกเสียงเป็น 2 พยางค์โดยพยางค์แรกเหมือนมีเสียงสระ อะ ประสมอยู่ เช่นคำว่า ขนม ออกเสียง ขะ-นม, แกลง ออกเสียงเป็น ละ-หล่ง เป็นต้น

จากการศึกษาส่วนประกอบต่างๆของคำ ทั้งในเรื่องหน้าที่ วิธีใช้งานและการออกเสียงดังที่ได้นำเสนอข้างต้น ผู้วิจัยได้ออกแบบกฎและตารางแปลงคำรหัสชาวดิถีแบบสมบูรณ์ โดยมีรายละเอียดดังต่อไปนี้

3.2 ออกแบบรหัสชาวดิถีแบบสมบูรณ์

รหัสชาวดิถีแบบสมบูรณ์เป็นรหัสที่นำเอาส่วนประกอบทั้งหมดของคำทั้งพยัญชนะต้น สระ (ทั้งรูปสระและความสั้นยาวของเสียงสระ) ตัวสะกด วรรณยุกต์ และตัวอักษรควบ มาใช้ในการแปลงคำรหัส โดยรหัสชาวดิถีของคำแต่ละคำประกอบด้วยส่วนย่อยเรียกว่า “รหัสพยางค์” ซึ่งจะมีเพียงหนึ่งรหัสพยางค์หรือหลายรหัสพยางค์ก็ได้ตามการออกเสียงของคำนั้นๆ แต่ละรหัสพยางค์มีความยาว 7 หลักและแบ่งออกเป็น 5 ส่วนย่อยได้แก่ รหัสพยัญชนะต้นคำ สระ ตัวสะกด วรรณยุกต์และตัวอักษรควบ ดังรายละเอียดแสดงในภาพที่ 2



ภาพที่ 2 ส่วนประกอบต่างๆ ของรหัสพยางค์ที่แปลงรหัสมาจากคำว่า “พลัง”

กฎการแปลงรหัสชาวเด็กซ์แบบสมบูรณ์

การแปลงรหัสถูกออกแบบและสร้างขึ้นตามหลักอักขรวิธีของการสะกดคำภาษาไทย ซึ่งสามารถสรุปเป็นกฎพื้นฐานและกฎที่เกี่ยวกับการแปลงรหัสส่วนต่างๆ อีก 4 ส่วนได้แก่พยัญชนะต้นและตัวควบ สระ ตัวสะกด และวรรณยุกต์ ดังรายละเอียดต่อไปนี้

3.2.1 กฎพื้นฐาน

1) ก่อนการแปลงรหัสให้ตัดเครื่องหมายวรรคตอน และตัวอักษรที่ตัวการันต์กำกับอยู่ทั้งกรณีที่เป็นตัวอักษรเดี่ยว ตัวอักษรควบ หรือตัวอักษรควบกับสระเช่น วงศ์ ตัดตัวอักษร ์, จันทรื ตัดตัวอักษร ทรื, ศักดิ์ ตัดตัวอักษร ดี, พันธุ์ ตัดตัวอักษร ธุ์ เป็นต้น

2) ทำการแปลงรหัสตัวอักษรที่เหลืออยู่ ตามกฎที่เกี่ยวข้องกับส่วนประกอบแต่ละส่วนของคำ ดังนี้

3.2.2 กฎการแปลงรหัส พยัญชนะต้นคำ

1) พยัญชนะต้นคำอาจเป็นตัวอักษรเดี่ยวหรือเป็นตัวอักษรควบก็ได้

2) กรณีที่เป็นตัวอักษรเดี่ยวแบ่งออกเป็น 3 กรณีคือ

(2.1) เป็นตัวอักษรแรกของคำ และตามด้วยสระหน้า (ได้แก่ เ, แ, ไ, ใ หรือ โอ) เช่นคำว่า สไบ, สเกต ให้ใช้ตัวอักษรตัวแรกเป็นพยัญชนะต้นโดยแปลงรหัสตามตารางที่ 5.1 (และกำหนดเสียงสระของพยางค์เป็น “ะ”)

(2.2) เป็นตัวอักษรแรกของคำ และตามด้วยสระอื่น เช่นคำว่า ปาก, ชัก, สูง ให้แปลงรหัสตัวอักษรนั้นตามตารางที่ 5.1 (และกำหนดเสียงสระของพยางค์ตามสระที่เป็นตัวอักษรตัวที่ 2)

(2.3) เป็นตัวอักษรเดี่ยวที่อยู่ตามหลังสระหน้า เช่นคำว่า ใจ, ไป, โท ให้แปลงรหัสตัวอักษรตามตารางที่ 5.1 (และกำหนดเสียงสระของพยางค์ตามสระที่เป็นตัวแรก)

3) กรณีที่เป็นตัวอักษรควบกับตัวอักษร ร, ล หรือ ว สามารถแบ่งออกได้เป็น 7 กรณี คือ

(3.1) หากเป็นตัวอักษรควบ ทร และตามด้วยสระหรือวรรณยุกต์ให้เปลี่ยน ทร เป็นตัวอักษร ช และแปลงรหัสตามตารางที่ 5.1 โดยไม่มีตัวอักษรควบ ตัวอย่างเช่น ทราบ เปลี่ยนเป็น ชาบ, ทรุด เปลี่ยนเป็น ชุด เป็นต้น

(3.2) หากเป็นตัวอักษรควบ ทล หรือ ทว และตามด้วยสระหรือวรรณยุกต์ เช่นคำว่า ทลาย, ทวี, ทวาร ให้แยกตัวอักษรควบออกเป็นพยัญชนะต้นของสองพยางค์โดยไม่มี

ตารางที่ 5.1 พยัญชนะและรหัสอักษรสำหรับพยัญชนะต้นคำของชาวเด็กซ์แบบสมบูรณ์

กลุ่มพยัญชนะต้นคำ	อักษรแทนกลุ่ม
ก	กก
ค, ค, ฅ	คค
ข, ข	คข
ง	งง
จ	จจ
ช, ฌ	ชช
ฉ	ชฉ
ซ, ฌร	ซซ
ศ, ษ, ส	ซศ
ด, ฎ	ดด
ต, ฏ	ตต
น, ณ	นน
ท, ฑ, ฒ	ทท
ฐ, ฎ	ทฐ
บ	บบ
ป	ปป
พ, ภ	พพ
ผ	พผ
ฟ	พฟ
ฝ	พฝ
ม	มม
ย, ญ	ยย
ร, ล, พ, ฤ, ฌ	รร
ว	วว
ฮ	ฮฮ
ห	ฮห
อ	ออ

ตัวอักษรควบ (พยางค์แรกออกเสียงสระ “ะ” และพยางค์ที่สองออกเสียงสระที่ปรากฏอยู่ในคำ เช่น ทลาย แยกเป็น ทะ-ลาย)

(3.3) หากตัวอักษรตัวแรกไม่ใช่ ท และตัวอักษรตัวที่ 3 เป็น สระหรือวรรณยุกต์ เช่นคำว่า กว้าง, กลาง, พระ, ขวัญ ให้ใช้ตัวอักษรตัวแรกเป็นพยัญชนะต้น และกำหนดให้ตัวอักษร ร, ล หรือ ว นั้นเป็นตัวอักษรควบ

(3.4) หากตัวอักษรที่สอง เป็น ร หรือ ล และตัวอักษรตัวที่ 3 เป็นตัวสะกด เช่นคำว่า ตรง, กลม, ปลด ฯลฯ ให้ใช้ตัวอักษรแรกเป็นพยัญชนะต้น และให้ตัวอักษร ร หรือ ล นั้นเป็นตัวตัวอักษรควบ (กำหนดเสียงสระของพยางค์เป็น “โอะ” และใช้ตัวอักษรตัวที่ 3 เป็นตัวสะกด)

ตารางที่ 5.2 การจัดกลุ่มสระและคำรหัสสำหรับสระของ
ชาวดี้เก็กซ์แบบสมบูรณ์

สระ	รหัสสระ	รหัสพยัญชนะต้น	รหัสตัวสะกด
ะ	1A		
า	1B		
อ	2C		
เ	2D		
เอ	3E		
แอ, อย	3F		
ง	4G		
น	4H		
เอ, เ	5I		
เ	5J		
แอ, แ	6K		
แ	6L		
โ	7M		
โ	7N		
เอา, เ	8O		
อ	8P		
เอา	9Q		
เอา	9R		
เอา, เ	AS		
เ, เ	AT		
เอา, เ	BU		
เอ, เ	BV		
ัว	CW		
ัว, ว	CX		
ฤ	3E	ร	
ຶ	3F	ร	
ຶ	3E	ล	
ຶ	3F	ล	
ำ	1A		ม
เ, เ	1A		ย
เ	1A		ว
รร	1A		น

3.2.4 กฎการแปลงรหัสตัวสะกด

1) กรณีที่ตัวสะกดเป็นอักษรเดี่ยว ให้พิจารณาดังนี้

(1.1) หากเป็นตัวอักษร a และมีสระ e อยู่ก่อนหน้า
ให้กำหนด a เป็นสระ

(1.2) หากเป็นตัวอักษร ย และมีสระ ใ อยู่ก่อนหน้า
ให้กำหนด ย เป็นสระ

(1.3) หากเป็นตัวอักษร w และมีสระ อยู่ก่อนหน้า
ให้กำหนด w เป็นสระ

(1.4) หากเป็นตัวอักษรเดียวกัน ให้แปลงรหัสตาม
ตารางที่ 5.3

2) กรณีที่ตัวสะกดเป็นตัวอักษรควบกับ ร หรือ ห ให้พิจารณา ดังนี้

(2.1) ตัวอักษรควม รค, กร คร ตร ทร หรือ ปร
ปรากฏอยู่หลังสระหรือวรรณยุกต์ เช่นคำว่า มารค, จักร,
สมัคร, บัตร, สมุทร, กอปร เป็นต้น ให้ทั้งตัวอักษร ร และใช้
ตัวอักษรที่เหลือเป็นตัวสะกด โดยแปลงรหัสตามตารางที่ 5.3

(2.2) ตัวอักษรควบ หม เช่นคำว่า พรหม, พรหมณ์
ให้ทั้งตัวอักษร ห และ ใ มี เป็นตัวสะกด

ตารางที่ 5.3 การจัดกลุ่มอักษรและคำรหัสสำหรับตัวสะกดของ
ชาวใต้เด็กซ์แบบสมบูรณ

กลุ่มอักษร	รหัส
ก, ข, ข, ค, ค, ง, จ, ช, ช, ฉ, กว, กร	ก
ด, จ, ช, ช, ฎ, ฏ, ฐ, ฑ, ฒ, ด, ต, ถ, ท, ฒ, ษ, ษ, ฒ, ฒ, ฒ, ทร, ทร	ด
บ, ป, ภา, พ, ฟ, ปร	บ
ง	ง
น, ญ, ณ, ร, ล, พ	น
ม, ำ	ม
ย, ใ, ใ	ย
ว, ะ	ว

3.2.5 กฎการแปลงรหัสวรรณยุกต์

การแปลงรหัสวรรณยุกต์ ให้แปลงรหัสตามค่าใน
ตารางที่ 5.4

ตารางที่ 5.4 ค่ารหัสของรูปวรรณยุกต์

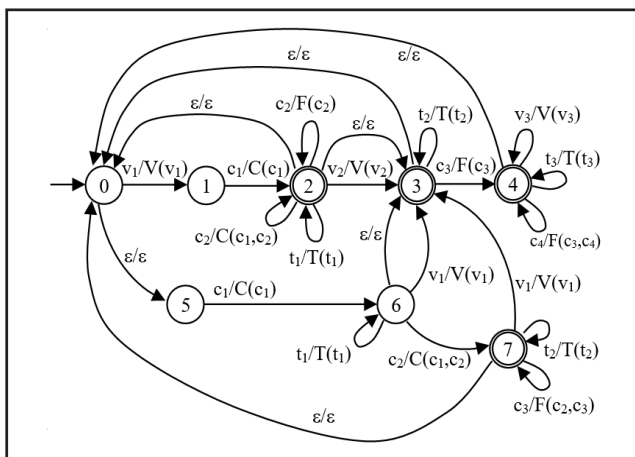
รูปวรรณยุกต์	ไม่มีรูป	ˊ	ˋ	ˊˋ	ˊˋˋ
คำรหัส	0	1	2	3	4



จากกฎการแปลงรหัสส่วนต่างๆ ดังที่กล่าวมาข้างต้น สามารถนำมาวิเคราะห์เพื่อใช้แยกส่วนประกอบต่างๆ ของคำ ดังตัวอย่างในตารางที่ 6

ตารางที่ 6 ตัวอย่างการวิเคราะห์ส่วนประกอบของคำตามกฎการแปลงรหัสพยัญชนะต้นคำ

คำ	พยางค์ ที่	พยัญชนะ ต้น	สระ	ตัว สะกด	วรรณยุกต์	อักษร ควบ
สไป	1	ส	ะ			
	2	บ	ไ			
กว้าง	1	ก	า	ง		ว
ทราบ	1	ท	า			
หลาย	1	ห	ะ			
	2	ล	า	ย		
กวาด	1	ก	ัว	ด		
อลเวง	1	อ	โ			
	2	ล	ะ			
	3	เ	ว			
ตรง	1	ด	โ	ง		ร
โลกา	1	ล	โ			
	2	ก	า			
ธนู	1	ธ	ะ			
	2	น	ุ			
อัตร	1	อ	ะ	ตร		
	2	ต	า			ร



ภาพที่ 3 โครงสร้าง Nondeterministic Finite Automaton แสดงกระบวนการแปลงรหัสรหัสชาวดิจิทัลแบบสมบรูณ์

3.2.6 กระบวนการแปลงรหัส

จากกฎต่างๆ ที่ได้กำหนดขึ้น สามารถนำมาเขียนกระบวนการทำงานของเครื่องจักรแปลงรหัส (Converter machine) เป็นโครงสร้าง Nondeterministic Finite Automaton with output ดังภาพที่ 3

c1, c2, c3, c4 หมายถึง ตัวอักษรที่เป็นพยัญชนะ
v0, v1, v2, v3 หมายถึง ตัวอักษรที่เป็นสระ
t1, t2, t3 หมายถึง ตัวอักษรที่เป็นวรรณยุกต์
C() หมายถึง ฟังก์ชันที่ทำหน้าที่แปลงรหัสพยัญชนะต้น
V() หมายถึง ฟังก์ชันที่ทำหน้าที่แปลงรหัสสระ
F() หมายถึง ฟังก์ชันที่ทำหน้าที่แปลงรหัสตัวสะกด
เครื่องจักรแปลงรหัส จะทำการประมวลผลโดยพิจารณาตัวอักษรในคำทีละตัวอักษร และสร้างฟังก์ชันผลลัพธ์ที่ทำหน้าที่แปลงรหัสส่วนต่างๆ ได้แก่ ฟังก์ชันแปลงรหัสพยัญชนะต้นคำ C() ฟังก์ชันแปลงรหัสสระ V() และฟังก์ชันแปลงรหัสตัวสะกด F() ได้เป็นรหัสชาวดิจิทัลแบบสมบรูณ์ ดังตัวอย่างในตารางที่ 7

3.2.7 การเปรียบเทียบหาค่าความเหมือน

ค่าความเหมือนระหว่างคำสองคำ $S(X, Y)$ สามารถคำนวณได้จากสูตรต่อไปนี้

$$S(X, Y) = \begin{cases} \sum_{i=1}^n \frac{sim(Cx_i, Cy_i)}{n} \\ sim(Cx_i, Cy_i) = \begin{cases} 1 & \text{if } \exists Cx_i = C \\ 0 & \text{if } \forall Cx_i \neq C \end{cases} \\ 1 \leq i \leq n \\ n = \max(X_{len}, Y_{len}) \end{cases} \quad (1)$$

โดย X, Y หมายถึง คำที่นำมาเปรียบเทียบกัน

Cx_i, Cy_i หมายถึง รหัสหลักที่ i ของ X และ Y ตามลำดับ
 $sim(Cx_i, Cy_i)$ หมายถึง ค่าความเหมือน ของรหัสหลักที่ i ของคำที่ 1 และ 2 ตามลำดับ

X_{len}, Y_{len} หมายถึง ความยาวของรหัสของ X และ Y

n หมายถึง ความยาวของรหัสที่มากที่สุดของ X หรือ Y
ตัวอย่างการคำนวณและเปรียบเทียบค่าความเหมือนแสดงดังตารางที่ 8 และ 9

ตารางที่ 7 การจัดกลุ่มสระและคำรหัสสำหรับสระของ
ชาวดัชนีแบบสมบูรณ

คำ	การเปลี่ยน สถานะ	ฟังก์ชัน ผลลัพธ์	รหัส ชาวดัชนี
ก้าน	056634	C(ก)T(ั) V(ก)F(น)	กก1Bน2-
อัตร	056344073	C(อ)V(ั)F(ต) F(ตร)C(ต)V(ก)	ออ1Aต0- ตต1B-0-
มารค	05634	C(ม)V(ก)F(ร) F(รค)	มม1B-ก0-
สวรค์	056777	C(ส)C(สว) F(วร)F(รร)	ชศ1A-0- วว1Aน0-
กลับ	056734	C(ก)C(กล) V(ั)F(บ)	กก1Aบ0ล
สามารถ	05634407344	C(ส)V(ก)F(ม)V(ก) C(ม)F(ร)F(รค)	ชศ1B-0- มม1Bต0-
กมล	05677	C(ก)C(กม) F(มล)	กก1A-0- มม7Mน0-
กมล	056773	C(ก)C(กม) F(มล)C(ล)V(ก)	กก1A-0- มม1A-0- รร1B-0-
ใกล้	01222	V(ิ)C(ก)C(กล)T(ั)	กก1Aย2ล
โง่ง	01222	V(ิ)C(ก)T(ั) F(ง)	กก7Nง1-
เครื่อง	01223344	V(ิ)C(ค)C(คร) V(ั)T(ั) F(อ)F(อง)	คคBVง1ร
กัม	05634	C(ก)T(ั)F(ม)	กก7Mม2-
แกน	0122	V(ิ)C(ก)F(น)	กก6Lน0-
อำนาจ	056305634	C(อ)V(ั) C(น)V(ก)F(จ)	ออ1Aม0- นน1Bต0-
ทราย	056734	C(ท)C(ทร)V(ก)F(ย)	ชช1Bย0-

ตารางที่ 8 ตัวอย่างการคำนวณหาความเหมือนระหว่าง
คำว่า ชัมชืน และ ชมชืน

คำ	รหัสชาวดัชนี													
ชัมชืน	ค	ช	7	M	2	ม	0	ค	ช	3	F	1	น	0
ชมชืน	ค	ช	7	M	0	ม	0	ค	ช	3	F	1	น	0
ความเหมือน	1	1	1	1	0	1	1	1	1	1	1	1	1	1

ตารางที่ 9 ตัวอย่างการคำนวณหาความเหมือนระหว่าง
คำต่าง ๆ ที่มีการสะกดคล้ายกัน

คำ		ชัมชืน	ชมชืน	ชบชืน
	รหัสชาวดัชนี	คช7M2 ม0คช 3F1น0	คช7M0 ม0คช 3F0น0	คช7M1 บ0คช 3F0น0
ชัมชืน	คช7M0ม0คช3F1น0	0.93*	0.93*	0.79
ชมชืน	คช7M1ม0คช3F0น0	0.86	0.93*	0.93*
ชืนชัม	คช3F0น0คช7M1ม0	0.50	0.50	0.43
ชัมชเง	คช7M1ม0ชห5J0ง0	0.50	0.57	0.57
ชัมชน	คช4G0ม0คช7M0น0	0.57	0.71	0.57
ชืนชืน	คช4G1น0คช7M2น0	0.50	0.50	0.57

3.2.8 ผลกระทบด้านเวลาที่ใช้ในการทำงาน

แม้ว่าการแปลงรหัสรหัสชาวดัชนีแบบสมบูรณ จะ
มีจำนวนกฎที่ใช้ในการแปลงรหัสมากกว่าการแปลงรหัส
แบบอื่นๆ แต่เวลาที่ใช้ในการแปลงรหัสส่วนใหญ่ คือ
การแปลงรหัสคำในพจนานุกรมซึ่งจะถูกดำเนินการไว้ก่อน
ที่จะมีการเปรียบเทียบหาความเหมือน ดังนั้นเมื่อมีการใช้งาน
จริงในขั้นตอนการเปรียบเทียบหาความเหมือน จึงคงเหลือ
เฉพาะการแปลงรหัสของคำที่พิมพ์ผิดเพียงคำเดียวเท่านั้น
ซึ่งใช้เวลาค่อนข้างน้อย จนไม่สามารถสังเกตความแตกต่าง
จากวิธีการแปลงรหัสแบบอื่นๆ ได้

3.3 การทดสอบและประเมินประสิทธิภาพ

ในการทดสอบ ผู้วิจัยได้กำหนดคำเป้าหมายทั้งที่เป็น
ชื่อคน และคำศัพท์ที่ใช้กันทั่วไปจำนวน 50 คำ ที่มีรูปแบบ
การสะกดต่างๆ ทั้งการใช้ตัวอักษรควบ สระหน้า การใช้ รร
ตัวการันต์ และการสะกดแบบอื่นๆ แล้วให้กลุ่มตัวอย่าง
จำนวน 30 คน เขียนคำสะกดผิดที่อาจเกิดขึ้นในลักษณะ
ต่างๆ จากคำเหล่านั้นทั้งสิ้น 1,014 คำ ซึ่งสามารถสรุปได้
ทั้งหมด 6 รูปแบบได้แก่ การพิมพ์ตัวอักษรไม่ครบ พิมพ์เกิน
พิมพ์ซ้ำ พิมพ์ผิด พิมพ์ตัวอักษรสลับที่และพิมพ์สะกด
แตกต่างออกไป ดังตัวอย่างบางส่วนที่แสดงในตารางที่ 10

จากนั้น จึงนำแต่ละคำมาแปลงเป็นรหัสชาวดัชนีแบบ
สมบูรณ เพื่อค้นคืนคำถูกต้องที่มีเสียงอ่านคล้ายกันจาก
พจนานุกรม



ตารางที่ 10 ตัวอย่างคำที่ใช้เป็นข้อมูลในการทดสอบ

คำถูก	คำผิดที่ใช้ในการทดสอบ					
	พิมพ์ไม่ครบ	พิมพ์เกิน	พิมพ์ซ้ำ	พิมพ์ผิด	พิมพ์สลับที่	สะกดแตกต่าง
ควาน	ควา, ควัน, วาน	คว้านล, คว้าน, กควาน	คว้านน, คควาน	คว้าง, คว้าน, ควัน	คว้านา, คว้าน, ควัน	คว้าน, คว้าน
ทราย	ทรา, ทรย	ทรายบ, ทรายา	ทรายย, ททราย, ทรราย	ทริย, ยราย, ทราง	ทารย, ทรายา, ทราย	ชาย
ตรง	ตร, รง, ดง	ตรงย, ตรง, ตยง	ตตรง, ตรงง, ตรงง	มรง, ตรง	ตงร, รตง	ตง
สไป	ไป, สไป	สโตไป, สไปค, สไปบ	สไปบ, สสไป	สไป, จไป, สไป	สไปบ, สไป, บไป	สไป, สบัย
ไกล	ไก, ไล, กล	ไกลี, ไกลน	ไกลล, ไกกล	เกล, ไกล, ไกล	กไล, ไกล	ไกล, ไกร, กลัย, กัย
วรรค	วรร, วรรค, วรร	วรรคส, วรรคค	วรรคค, วรรคค, วรรคค	วรรค, วรรค, วรรค	วรรค, วรรค, วรรค	วัค, วก, วัค
จรรยา	จรรยา, จรรา, รรยา	จรรยาย, จรรยสา	จจรรยา, จจรรยา			จันยา, จัญญา
อาจารย์	อาจารย์, อาจารย์	อาจารย์, อวอาจารย์	อาจารย์, อาจารย์	นาวารย์, อูจารย์, อาจารย์	จาจารย์, อายาร์จ	อาจาน, อาจาน
พัคค์	พัคค์, พัคค์, พัคค์	พัคค์, พัคค์, พัคค์	พัคค์, พัคค์, พัคค์	พัคค์, พัคค์, พัคค์	พัคค์, พัคค์	พัค, ภัค
อลเวง	อลเวง, ลเวง, อลเวง	อัลเวง, อลเวงา, อลเวง	อลเวง, อลลเวง, อลเวง	อนเวง, อลเวง, สลเวง	ลอลเวง, อลเวง, อลเวง	อนล, ะเวง, โอน, ละเวง
อย่า	อย่า, อย่า, อย่า	อย่า, อย่า, อย่า	อย่า, อย่า, อย่า	อย่า, กยา, ฮยา	อย่า, อย่า	หยา
ขมขึ้น	ขขึ้น, มขึ้น, ขมขึ้น	ขมขึ้น, ขมขึ้น, ขมขึ้น	ขมขึ้น, ขมขึ้น, ขมขึ้น	ขขึ้น, ขมขึ้น, ขมขึ้น	มขขึ้น, ขมขึ้น	ขมขึ้น, ขมขึ้น
อุไรวรรณ	อุไรวรรณ, อไรวรรณ	อุไรวรรณ, อไรวรรณ, อไรวรรณ	อุไรวรรณ, อไรวรรณ, อไรวรรณ	อุไรวรรณ, อไรวรรณ, อไรวรรณ	อุไรวรรณ, อไรวรรณ	อุไรวัน, อุไรวัน
ดวงหทัย	ดงหทัย, วงหทัย, ดงหทัย	ดาวหทัย, ดงนหทัย	ดงนหทัย, ดงนหทัย	กวงหทัย, ดาวหทัย, ดงนหทัย	ดงนหทัย, ดงนหทัย	ดง, หะไท, ดง, หะทัย
อรอนงค์	อรอนงค์, อรอนงค์	อรอนงค์, สอรอนงค์	อรอนงค์, อรอนงค์	อรอนงค์, อรอนงค์, สอรอนงค์	อรอนงค์, อรอนงค์, อรอนงค์	อรอนงค์, ออรอนงค์

4. ผลการดำเนินงาน

ผลการทดสอบประสิทธิภาพของรหัสชาวเด็กซ์แบบสมบูรณ์ พบว่าจากคำผิดที่ทดสอบค้นคืนทั้งหมด 1,014 คำ สามารถค้นคืนคำที่ถูกต้อง (True Positive) ได้จำนวน 951 คำ โดยมีการค้นคืนคำผิดพลาด (False Positive) จำนวน 80 คำ และมีคำที่ถูกที่ไม่ถูกค้นคืน (False Negative) จำนวน 63 คำ ดังตารางที่ 11

ตารางที่ 11 ผลทดสอบการค้นคืนคำด้วยวิธีแปลงรหัสชาวเด็กซ์แบบสมบูรณ์

ลักษณะของคำผิด	ประเภทของคำ	จำนวนของคำ			
		คำผิดที่ใช้ทดสอบ	ค้นคืนถูกต้อง (True Positive)	ค้นคืนผิดพลาด (False Positive)	คำถูกที่ไม่ถูกค้นคืน (False Negative)
พิมพ์ไม่ครบ	มีตัวควบ	28	25	7	3
	ใช้สระหน้า	26	23	4	3
	มี รร	28	24	5	4
	มีการันต์	30	26	2	4
	แบบอื่นๆ	31	29	3	2
	ชื่อบุคคล	35	32	2	3
	รวม	178	159	23	19
พิมพ์เกิน	มีตัวควบ	29	26	4	3
	ใช้สระหน้า	30	28	3	2
	มี รร	28	25	2	3
	มีการันต์	28	26	3	2
	แบบอื่นๆ	29	26	3	3
	ชื่อบุคคล	36	33	3	3
	รวม	180	164	18	16
พิมพ์ซ้ำ	มีตัวควบ	30	29	1	1
	ใช้สระหน้า	29	29	0	0
	มี รร	28	27	2	1
	มีการันต์	29	28	1	1
	แบบอื่นๆ	28	26	2	2
	ชื่อบุคคล	35	35	2	0
	รวม	179	174	8	5



ตารางที่ 11 ต่อ

ลักษณะของคำผิด	ประเภทของคำ	จำนวนของคำ			
		คำผิดที่ใช้ทดสอบ	ค้นคืนถูกต้อง (True Positive)	ค้นคืนผิดพลาด (False Positive)	คำถูกที่ไม่ถูกค้นคืน (False Negative)
พิมพ์ผิด	มีตัวควบ	30	27	2	3
	ใช้สระหน้า	31	29	3	2
	มี รร	27	26	3	1
	มีการันต์	29	28	2	1
	แบบอื่นๆ	29	27	4	2
	ชื่อบุคคล	37	35	1	2
	รวม	183	172	15	11
พิมพ์สลับที่	มีตัวควบ	26	25	0	1
	ใช้สระหน้า	29	29	2	0
	มี รร	28	27	1	1
	มีการันต์	30	29	2	1
	แบบอื่นๆ	31	28	2	3
	ชื่อบุคคล	38	37	3	1
	รวม	182	175	10	7
สะกดแตกต่าง	มีตัวควบ	19	18	1	1
	ใช้สระหน้า	18	16	0	2
	มี รร	18	17	1	1
	มีการันต์	20	19	1	1
	แบบอื่นๆ	17	17	2	0
	ชื่อบุคคล	20	20	1	0
	รวม	112	107	6	5
ยอดรวมทั้งหมด		1014	951	80	63

ผลการทดสอบพบว่ามีความแม่นยำ (Precision) การเรียกคืน (Recall) และคะแนนเอฟวัน (F1 score) ที่ 0.92, 0.94 และ 0.93 ตามลำดับ

นอกจากนี้ เมื่อนำผลการทดลองไปเปรียบเทียบกับวิธีการแปลงรหัสด้วยวิธีการอื่นๆ ที่มีมาก่อนหน้านี้ 3 วิธี พบว่าการแปลงรหัสชาวดึกซ์แบบสมบูรณ เป็นวิธีการที่สามารถวิเคราะห์หาคำที่ออกเสียงคล้ายกันได้ดีกว่า โดยสามารถค้นหาคำที่วิธีการอื่นๆ ไม่สามารถค้นคืนได้พร้อมทั้งระบุค่าระดับความเหมือนของแต่ละคำที่ค้นคืนได้ด้วยดังตัวอย่างในตารางที่ 12

ตารางที่ 12 เปรียบเทียบรหัสชาวดึกซ์และค่าความเหมือนของการค้นคืนคำสะกดผิด “ปัญญา” ด้วยวิธีชาวดึกซ์แบบต่างๆ

รหัส และค่าความเหมือนระหว่างคำ ปัญญา กับคำอื่น ๆ				
คำ	แบบสมบูรณ	วิจิต	วรรณิ	เทพพิทักษ์
ปัญญา	ปป4G0น- ยย1B0--*	ปE490*	ป330000*	ปนนยาส, ปนนยาสยาส
ปัญญา	ปป1A0น- ยย1B0--*	ป4900	ป330000*	ปานยาส, ปานยาสยาส
	0.86	0	1.00	0
บุญญา	บบ4G0น- ยย1B0--*	บE490	บ330000	บนนยาส, บนนยอนยาส, บนนยาสยาส, บนนยคยาส, บนนยาสยาส
	0.86	0	0	0
บุณญา	บบ4G0น- ยย1B0--*	บE490	บ320000	บนนยาส, บนนยาสยาส, บนนยคยาส, บนนยาสยาส
	0.86	0	0	0
ปันนา	ปป1A0น- นน1B0--	ป4900	ป330000*	ปานนยาส, ปานนยาสยาส
	0.71	0	1.00	0

จากตารางจะเห็นว่า ชาวดึกซ์แบบสมบูรณ สามารถค้นหาคำถูกต้องที่ออกเสียงคล้ายกับคำว่า “ปัญญา” ได้ดีกว่าวิธีอื่นๆ โดยเป็นคำที่มีค่าความเหมือนสูงกว่า 0.8 จำนวน 3 คำได้แก่ “ปัญญา”, “บุญญา” และ “บุณญา” ในขณะที่วิธีการชาวดึกซ์ของวิจิตและเทพพิทักษ์ ไม่พบคำที่มีรหัสชาวดึกซ์ตรงกัน ส่วนวิธีของวรรณิ พบคำที่มีรหัสชาวดึกซ์ตรงกันสองคำคือคำว่า “ปัญญา” และ “ปันนา” ทั้งที่คำว่า “ปันนา” นั้นออกเสียงแตกต่างจากคำว่า “ปัญญา” ก่อนข้างมาก ซึ่งผลการทดสอบค้นหาด้วยคำสะกดผิดในลักษณะอื่นๆ ทั้ง 6 รูปแบบ ก็ได้ผลออกมาในลักษณะเดียวกัน

5. สรุป

จากการทดสอบเปรียบเทียบประสิทธิภาพในการค้นคืนคำที่ออกเสียงเหมือนคำที่สะกดผิด ด้วยวิธีการแปลงรหัสแบบต่างๆ พบว่าการแปลงรหัสชาวดึกซ์แบบสมบูรณ เป็น

วิธีการที่มีประสิทธิภาพสูงที่สุด ในการวิเคราะห์หาคำที่ออกเสียงคล้ายกันจากชื่อและคำที่พิมพ์ผิดในลักษณะต่างๆ ทั้ง 6 รูปแบบ เนื่องจากเป็นวิธีที่ใช้ส่วนประกอบที่ทำให้เกิดเสียงอ่านของคำทั้งหมดในการเข้ารหัส และทำการค้นคืนคำที่มีค่าความเหมือนใกล้เคียงกัน โดยไม่ได้จำกัดอยู่เฉพาะการเปรียบเทียบคำที่มีรหัสตรงกันเท่านั้น ทำให้สามารถค้นคืนคำอื่นๆ ที่การแปลงรหัสด้วยวิธีต่างๆ ที่มีมาก่อนหน้านี้ไม่สามารถค้นได้เช่น คำที่พิมพ์ตัวอักษรขึ้นต้นคำผิด หรือคำที่มีความผิดพลาดในส่วนหนึ่งของ สระ วรรณยุกต์ อักษร “รร” และ ตัวควบทั้งในกรณีที่เป็นพยัญชนะต้นคำและตัวสะกด นอกจากนี้การแปลงรหัสชาวดึกซ์แบบสมบูรณ์ยังสามารถบอกถึงความแตกต่างของระดับความเหมือนระหว่างคำที่นำมาเปรียบเทียบกันได้ด้วย ซึ่งไม่มีในวิธีการแปลงรหัสแบบต่างๆ ก่อนหน้านี้

6. เอกสารอ้างอิง

- [1] National Archives and Records Administration. *The Soundex Indexing System*. Available Online at <https://www.archives.gov/research/census/soundex.html>. accessed on 24 December 2010.
- [2] Datamat Co. Ltd. *Local tax collection monitoring system*. Report to the Office of Policy and Planning, Bangkok, 1978.
- [3] C. Tapsai. “Searching Model on Learning Document by Soundex and Semantic Analysis.” *Proceedings of the AC-EMI-2018 International Academic Conference on Economy & Management Innovations*, Budapest, Hungary, pp. 159-163, 2018.
- [4] J. M. Chung, et al. “Automatic English-Chinese name translation by using Web-Mining and phonetic similarity.” *Paper presented at the 2011 IEEE International Conference on Information Reuse & Integration*, 2011.
- [5] T. Karoonboonyanan and et al. “A Thai Soundex System for Spelling Correction”. *Proceedings of NLPRS*, pp. 633-644, 1997.
- [6] D. Li, and D. Peng. “Spelling Correction for Chinese Language Based on Pinyin-Soundex Algorithm.” *Proceedings of International Conference on Internet Technology and Applications*, 2011.
- [7] N. Angkawattanawit, C. Haruechaiyasak, and S. Marukatat. “Thai Q-Cor: Integrating Word Approximation and Soundex for Thai Query Correction.” *Proceedings of International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology*, Krabi, Thailand, pp.121-124, 2008.
- [8] V. Lorchirachoonkul. “A Thai soundex system.” *Information Processing & Management*, Vol. 18, No. 5, pp. 243-255, 1982.
- [9] W. Udompanich. *String searching for Thai alphabet using Soundex compression technique*. Master Thesis of Department of Computer Engineering Graduate School, Chulalongkorn University, 1983.
- [10] U. Silapasarn. *Thai Language Principle*. Thai Wattanapanich Publishing, Bangkok, Thailand, pp. 1-26, 1990.
- [11] K. Thonglo. *Thai Language Principle*. Ruamsarn co.,ltd., Bangkok, Thailand, pp. 54-131, 1994.