หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีต่าง ๆ ที่เกี่ยวกับหลักภาษาไทย ได้แก่ การถอดอักษร การใช้อักษรโรมันเพื่อการถ่ายเสียง และหลักเกณฑ์การทับศัพท์ เพื่อใช้เป็นฐานความรู้ในการออกแบบขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษให้ทำงานได้ถูกต้องตามหลักภาษา

ส่วนงานวิจัยต่าง ๆ ที่ผ่านมาที่เกี่ยวข้องและมีอิทธิพลต่องานวิจัยนี้ได้แก่ การค้นคืนสารสนเทศข้ามภาษา ขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษ ขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษ และขั้นตอนวิธีระยะแก้ไขสั้นที่สุด

2.1 การถอดอักษร

การถอดอักษร (Transliteration) หมายถึง การนำคำในภาษาหนึ่งมาเขียนด้วยตัวอักษรอีกภาษาหนึ่งแบบอักษรต่ออักษร โดยพยายามใช้หน่วยเสียงของอักษรทั้งสองภาษาใกล้เคียงกันมากที่สุด [1] ตัวอย่างเช่น คำว่า "HARVARD" ในภาษาอังกฤษถอดอักษรเป็น "ฮาร์วาร์ด" ในภาษาไทย เป็นต้น การถอดอักษรแบ่งเป็น 3 ขั้นตอนหลัก ๆ ดังนี้

1. ถอดหน่วยอักษรในภาษาต้นแบบ (Source Language) เป็นหน่วยเสียงในภาษาต้นแบบ เช่น ถอดหน่วยอักษร "B" ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาอังกฤษ เป็นต้น
2. แทนหน่วยเสียงในภาษาต้นแบบ ด้วยหน่วยเสียงในภาษาเป้าหมาย (Target Language) โดยพยายามใช้หน่วยเสียงที่ใกล้เคียงกันมากที่สุด เช่น แทนหน่วยเสียง /b/ ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาไทย เป็นต้น
3. ถอดหน่วยเสียงในภาษาเป้าหมาย เป็นหน่วยอักษรในภาษาเป้าหมาย เช่น ถอดหน่วยเสียง /b/ ในภาษาไทยเป็นหน่วยอักษร "บ" ในภาษาไทย เป็นต้น

---

[1] อุไรรัตน์ บุญภานนท์, "การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักวิชาภาษาศาสตร์" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2528), หน้า 6-18.

จากขั้นตอนดังกล่าวสามารถแสดงตัวอย่างการถอดอักษรได้ดังตารางที่ 2.1

ภาษาต้นแบบ		ภาษาเป้าหมาย	
หน่วยอักษร	หน่วยเสียง	หน่วยเสียง	หน่วยอักษร
B	/b/	/b/	บ
M	/m/	/m/	ม
T	/t/	/th/	ท

ตารางที่ 2.1 ตัวอย่างการถอดอักษรจากภาษาต้นแบบไปยังภาษาเป้าหมาย

ปัญหาต่าง ๆ ในการถอดอักษร

- ความสัมพันธ์ของหน่วยอักษรและหน่วยเสียง มีความสัมพันธ์แบบหนึ่งตัวอักษรแทนหลายหน่วยเสียง เช่น ในภาษาอังกฤษ "C" แทนด้วย /k/ หรือ /s/ เป็นต้น และมีความสัมพันธ์แบบหลายหน่วยอักษรแทนหนึ่งหน่วยเสียง เช่น ในภาษาอังกฤษ "N, TN, GN, PN" แทนด้วย /n/ ในภาษาไทย "ร, ฤ, หร" แทนด้วย /r/ และ "ฉ, ช, ฌ" แทนด้วย /ch/ เป็นต้น
- การแบ่งพยางค์ในภาษาต้นแบบ เมื่อมีพยัญชนะตัวเดียวอยู่ระหว่างสระ เช่น คำว่า money ในภาษาอังกฤษ จะแบ่งพยางค์อย่างไร จะถอดพยัญชนะซ้ำสองตัวเพื่อให้อ่านได้สะดวกเป็น มัน-นีย์ หรือจะถอดอักษรเพียงตัวเดียวตามที่ปรากฏในภาษาอังกฤษเป็น มะ-นีย์ หรือ มัน-อีย์
- ปัญหาอันเนื่องมาจากช่วงเวลาของการยืมคำทับศัพท์ คำทับศัพท์บางคำยืมมาเป็นเวลานาน ซึ่งในอดีตมีหลักเกณฑ์การทับศัพท์ไม่ตรงกับหลักเกณฑ์ในปัจจุบัน เช่น "C" ที่แทน /k/ ในอดีตนิยมถอดเป็นอักษร "ก" เช่น กุ๊ก (Cook) กัปตัน (Captain) กะรัต(Carat) แก๊ป(Cap) เป็นต้น แต่ปัจจุบัน "C" ที่แทน /k/ มักจะถอดเป็น "ค" ในตำแหน่งพยัญชนะต้น เช่น คอนโดมิเนียม (Condominium) แคปซูล (Capsule) แครอต (Carrot) เป็นต้น

2.2 การใช้ตัวอักษรโรมันเพื่อการถ่ายเสียง

การใช้ตัวอักษรโรมันเพื่อการถ่ายเสียง (Romanization) [2] คือการถ่ายเสียงตัวอักษรของภาษาอื่น ๆ ที่ไม่ใช่อักษรโรมัน เช่น ไทย จีน ญี่ปุ่น ฯลฯ ให้เป็นตัวอักษรโรมัน เพื่อให้ผู้ที่ไม่รู้จักภาษานั้น ๆ สามารถอ่านออกเสียงได้ เช่น คำว่า "อยุธยา" สามารถถ่ายเสียงเป็น "AYUTTHAYA" เป็นต้น การเขียนชื่อเฉพาะต่าง ๆ จากภาษาไทยด้วยอักษรโรมันนั้นมีปัญหามาก เช่น คำว่า "ไทยทนุ" เขียนเป็น "Thai Danu" ซึ่งการใช้ตัวอักษร "D" แทนตัว "ท" นั้นเป็นการเขียนโดยอาศัยหลักเกณฑ์ของภาษาบาลี [3]

จากปัญหาต่าง ๆ ที่เกิดขึ้น ทางราชบัณฑิตยสถานจึงได้พยายามสร้างระบบมาตรฐานในการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงสำหรับตัวอักษรไทย โดยมีแนวคิดว่า ควรยึดหลักภาษาไทยเป็นที่ตั้งและหาวิธีการถ่ายเสียงตัวอักษรตามวิธีเขียนและการออกสำเนียงในภาษาไทย เพราะถ้ายึดหลักการถ่ายเสียงตัวอักษรตามศัพท์ในภาษาเดิมซึ่งเป็นที่มาของคำไทยแล้ว อักษรไทยตัวเดียวกันอาจถ่ายเสียงเป็นอักษรโรมันต่างกันก็ได้ เช่น ทร จะถ่ายเสียงเป็น Thon, S, Tr., Tara แต่ถ้าจะถ่ายเสียงตัวอักษรตามเสียงอย่างเดียวเท่านั้น เมื่อเขียนกลับเป็นภาษาไทยก็เขียนได้หลายอย่าง เช่น Ban Ma เขียนเป็น บ้านมา, บ้านม่า, บ้านม้า หรือ บ้านหมาได้ เป็นต้น

ในที่สุด ทางราชบัณฑิตยสถานจึงได้กำหนดระบบการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงสำหรับตัวอักษรไทยออกเป็น 2 ระบบ [4] คือระบบทั่วไปและระบบพิสดาร โดยระบบทั่วไปจะใช้สำหรับกรณีที่การออกเสียงสำคัญกว่าการเขียนตัวสะกด ซึ่งจะอาศัยหลักการออกเสียงเป็นสำคัญต้องสอดคล้องกับไวยากรณ์ของไทย และสามารถขยายเป็นระบบเฉพาะได้ เช่น คำว่า "กษัตริย์" ถ่ายเสียงเป็น Kasat ส่วนระบบเฉพาะจะใช้ในกรณีที่จะแสดงตัวอักษรให้ละเอียดแม่นยำ เพื่อให้คงความหมายของคำนั้นไว้ เช่น คำว่า "กษัตริย์" ถ่ายเสียงเป็น Kasatriy

---

[2] อุไรรัตน์ บุญภานนท์, "การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักวิชาภาษาศาสตร์" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2528), หน้า 6-18.

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

[4] อุไรรัตน์ บุญภานนท์, "การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักวิชาภาษาศาสตร์" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2528), หน้า 6-18.

2.3 หลักเกณฑ์การทับศัพท์

ในปัจจุบันได้มีการบัญญัติศัพท์วิชาการขึ้นใช้กันอย่างแพร่หลาย และในการบัญญัติศัพท์นั้น บางครั้งไม่สามารถหาคำไทยมาใช้ได้ตรงความหมายที่ต้องการ หรือกรณีที่เป็นคำที่ไม่เคยมีการบัญญัติศัพท์มาก่อนจึงต้องใช้วิธีทับศัพท์ นอกจากนี้การเขียนคำวิสามานยนามต่าง ๆ เช่น ชื่อคน ชื่อสถานที่ ก็ต้องใช้วิธีทับศัพท์เช่นเดียวกัน เพื่อให้เป็นมาตรฐานเดียวกันในการทับศัพท์ [5] ทางราชบัณฑิตยสถานจึงได้กำหนดหลักเกณฑ์การทับศัพท์ไว้ดังนี้

1. การทับศัพท์ให้ถอดอักษรในภาษาเดิมพอควรแก่การแสดงที่มาของรูปศัพท์ และให้เขียนในรูปที่อ่านได้สะดวกในภาษาไทย

2. การวางหลักเกณฑ์ได้แยกกำหนดหลักเกณฑ์การทับศัพท์ภาษาต่าง ๆ แต่ละภาษาไป เช่น หลักเกณฑ์การทับศัพท์สำหรับภาษาอังกฤษ หลักเกณฑ์การทับศัพท์สำหรับภาษาญี่ปุ่น เป็นต้น

3.คำทับศัพท์ที่ใช้กันมานานจนถือเป็นคำไทย และปรากฏในพจนานุกรมฉบับราชบัณฑิตยสถานแล้ว ให้ใช้ต่อไปตามเดิม เช่น ช็อกโกแลต ช็อกโกแลต เชื้ด แก๊ส ก๊าซ

4. ศัพท์วิชาการซึ่งใช้เฉพาะกลุ่มไม่ใช่ศัพท์ทั่วไปอาจเพิ่มหลักเกณฑ์ขึ้นตามความจำเป็น

2.4 การค้นคืนข้ามภาษา

การค้นคืนสารสนเทศข้ามภาษา หมายถึง การค้นคืนสารสนเทศ โดยภาษาที่ใช้ในข้อคำถามแตกต่างจากภาษาที่ใช้ในการจัดเก็บเอกสารแล้วระบบจะค้นคืนสารสนเทศข้ามภาษาให้โดยอัตโนมัติ โดยทั่วไปแล้วสามารถแบ่งการค้นคืนสารสนเทศข้ามภาษาออกเป็น 2 วิธี [6] คือ

1. การแปลข้อคำถาม คือ การแปลภาษาของข้อคำถามที่ผู้ใช้ระบุขณะที่ทำการสืบค้น ให้เป็นภาษาเดียวกับที่จัดเก็บในเอกสารก่อน แล้วจึงทำการสืบค้น

2. การแปลเอกสาร คือ การแปลภาษาของเอกสารทั้งหมดให้เป็นภาษาเดียวกับข้อคำถามก่อนแล้วจึง ทำการสร้างดัชนี

---

[5] ราชบัณฑิตยสถาน, หลักเกณฑ์การทับศัพท์ ฉบับราชบัณฑิตยสถาน (2535).

[6] D. Oard and B. Dorr, A Survey of Multilingual Text Retrieval, Technical Report UMIACS-TR-96-19 CD-TR-3615, University of Maryland, College Park, April 1996.

งานวิจัยทางด้านการค้นคืนสารสนเทศข้ามภาษาที่ผ่านมานั้น สามารถแบ่งออกเป็น 3 แนวคิดหลัก ๆ [7] คือ

1. การแปลข้อความโดยใช้เทคนิคการแปลภาษาด้วยเครื่อง (Machine Translation) [8] คือ การแปลภาษาที่เอกสารหรือข้อคำถามด้วยเครื่องแปลภาษา ซึ่งข้อด้อยของวิธีการนี้ขึ้นอยู่กับคุณภาพของการแปลภาษา

2. การใช้ฐานพจนานุกรมหลายภาษา (Multilingual Dictionary) [9] คือ การใช้พจนานุกรมหลายภาษาในลักษณะของบัญชีคำอรรถาภิธาน หรือการสร้างบัญชีคำพ้องความหมายแต่เป็นการพ้องความหมายข้ามภาษา เช่น คำว่า "Food" จะพ้องความหมายกับคำว่า "อาหาร" เป็นต้น

3. การใช้ฐานคำแบบขนาน (Parallel Corpus-Based) [10] เป็นอีกทางเลือกของการใช้บัญชีคำอรรถาภิธาน คือการใช้ประโยชน์จากฐานข้อมูลที่มีอยู่แล้ว ซึ่งจัดเก็บในลักษณะสองภาษาควบคู่กันไป ตัวอย่างเช่น ข้อมูลหนังสือที่มีการจัดเก็บชื่อเรื่องที่เป็นภาษาไทยและภาษาอังกฤษควบคู่กันไป เป็นต้น โดยจะนำข้อมูลแต่ละคู่มาจัดเก็บเป็นฐานคำแบบขนาน

2.5 ขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษ

M. K. Odell และ R. C. Russell ได้ออกแบบขั้นตอนวิธีการเข้ารหัสชื่อในภาษาอังกฤษ โดยยึดหลักของการอ่านออกเสียง เพื่อให้ชื่อที่อ่านออกเสียงคล้ายกันได้รหัสเหมือนกัน หรือที่เรียกว่า "ซาวด์เด็กซ์" (Soundex) ขั้นตอนวิธีดังกล่าวได้ใช้แนวคิดทางภาษาศาสตร์และชวเลขที่ว่าชื่อในภาษาอังกฤษสามารถจำแนกความแตกต่างได้โดยพิจารณาเพียงพยัญชนะเท่านั้น [11]

---

[7] D. Oard and B. Dorr, A Survey of Multilingual Text Retrieval, Technical Report UMIACS-TR-96-19 CD-TR-3615, University of Maryland, College Park, April 1996.

[8] B. R. Pevzner, Automatic Translation of English Text to the Language of the Pusto-Nepusto-2 System, Automatic Documentation and Mathematical Linguistics, 3(40):40-48, 1969.

[9] F. Semturs, STAIRS/TLS-A System for Free Text and Descriptor Searching, Proceedings of the ASIS Annual Meeting, Volume 15, pp. 295-298, November 1978.

[10] G. Salton, Experiments in Multi-Lingual Information Retrieval, Information Processing Letters, 2(1):6-11, 1973.

[11] A. Binstock and J. Rex, Practical Algorithms for Programmers (New York: Addison Wesley, 1995), pp. 157-172.

```
char *SOUNDEX(char *Name)
{
                    /*ABCDEFGHIJKLMNOPQRSTUVWXYZ*/
    char Table[] = "01230120022455012623010202";
    char Code[] = "0000";
    int   Count = 0;
    char Ch;
    /*----- For the First Character -----*/
    Code[Count++] = Name[0];
    /*----- For the Rest Character -----*/
    for (i=2; i < strlen (Name); i++) {
        Ch = Table[Name[i] - 'A'];
        if (Ch != PrevCode && Ch != '0') {
            Code[Count++] := Ch;
            if (Count = 5)
                return(Code);
        }
        PrevCode := Ch;
    }
    return(Code);
}
```

รูปที่ 2.1 โปรแกรมการเข้ารหัสซาวด์เด็กซ์ภาษาอังกฤษ

จากรูปที่ 2.1 แสดงขั้นตอนการเข้ารหัสซาวด์เด็กซ์ โดยเริ่มจากการนำตัวอักษรตัวแรกของคำไปเป็นรหัส ส่วนตัวอักษรที่เหลือจะแปลงเป็นตัวเลขโดยใช้ตารางการกำหนดรหัสซาวด์เด็กซ์ดังแสดงในตารางที่ 2.2 จากนั้นจะตัดรหัสตัวเลขศูนย์ออกไป และถ้ารหัสตัวเลขที่อยู่ตำแหน่งติดกันมีค่าเท่ากันจะเก็บเพียงหนึ่งรหัสเท่านั้น สุดท้ายรหัสซาวด์เด็กซ์ที่ได้คือตัวอักษรตัวแรกของชื่อตามด้วยรหัสตัวเลขสามตัวแรกที่ได้จากการแปลง ถ้ารหัสที่ได้มีความยาวไม่ถึงสี่หลักจะเติมตัวเลขศูนย์จนครบสี่หลัก ตัวอย่างเช่น คำว่า "ALEXANDER" มีรหัสซาวด์เด็กซ์เท่ากับ A425 เป็นต้น

ตัวอักษร	รหัสตัวเลข
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

ตารางที่ 2.2 การกำหนดรหัสซาวด์เด็กซ์ภาษาอังกฤษ

ขั้นตอนวิธีซาวด์เด็กซ์ดังกล่าวเป็นขั้นตอนวิธีที่ง่ายและทำงานได้รวดเร็ว ส่วนในเรื่องของประสิทธิผลในการค้นคืนพบว่าเปอร์เซ็นต์ของค่าเรียกคืนสูงมาก แต่เปอร์เซ็นต์ของค่าแม่นยำต่ำ

2.6 ขั้นตอนวิธีซาวด์เด็กซ์ภาษาไทย

ปัจจุบันมีงานวิจัยทางด้านซาวด์เด็กซ์ภาษาไทย เพื่อแก้ไขปัญหาต่าง ๆ เช่น การสืบค้นชื่อและนามสกุลที่อ่านออกเสียงเหมือนกันในทะเบียนรายชื่อ การสืบค้นคำไทยที่มักสะกดผิด และการตรวจสอบตัวสะกด เป็นต้น การออกแบบและพัฒนาขั้นตอนวิธีเพื่อแก้ปัญหาดังกล่าวจึงแตกต่างกันไป ในที่นี้จะขอกล่าวถึงงานวิจัยของวรรณี อุดมพานิชย์ [12] งานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา [13] และงานวิจัยของ Theppitak Karoonboonyanan [14]

---

[12] วรรณี อุดมพาณิชย์, "การใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์ มหาวิทยาลัย, 2526), หน้า 8-19.

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

[14] T. Karoomboonyanan, V. Somlertlamvanich and S. Meknavin, A Thai Soundex System for Spelling Correction, Proc. Of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, December 2-4, pp.633-636.

งานวิจัยของวรรณี อุดมพานิชย์ มีการออกแบบขั้นตอนวิธีการเข้ารหัสซาวด์เด็กซ์ภาษาไทยเพื่อให้ชื่อที่อ่านออกเสียงเหมือนกันได้รหัสตรงกัน เพื่อแก้ปัญหาในการสืบค้นชื่อและนามสกุลที่อ่านออกเสียงเหมือนกันในทะเบียนรายชื่อ หลักเกณฑ์และข้อกำหนดของขั้นตอนวิธีดัดแปลงมาจากซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russell ซึ่งได้เพิ่มเติมข้อกำหนดบางส่วนเพื่อเหมาะสมกับภาษาไทย หลักเกณฑ์การสร้างรหัสซาวด์เด็กซ์มีดังนี้

- กำหนดให้ขนาดของรหัสมีความยาวเท่ากับ 7 หลัก ประกอบด้วยตัวอักษร 1 หลักและตัวเลข 6 หลัก โดยมีรูปแบบของรหัสซาวด์เด็กซ์ดังนี้

  (รหัสซาวด์เด็กซ์) = (ตัวอักษร)(ตัวเลข)(ตัวเลข)(ตัวเลข)(ตัวเลข)(ตัวเลข)(ตัวเลข)
- เปลี่ยนพยัญชนะตัวแรกของคำเป็นรหัสตัวอักษรตามตารางที่ 2.3
- เปลี่ยนอักขระตั้งแต่ตัวที่สองขึ้นไปของคำเป็นรหัสตัวเลขตามตารางที่ 2.4
- โดยปกติจะไม่นำสระ วรรณยุกต์ และไม้ไต่คู้ มาสร้างรหัส ยกเว้นสระที่ให้เสียงสะกดเป็นพยัญชนะ ได้แก่ สระ -ำ ไ ใ
- ถ้ารหัสซาวด์เด็กซ์ที่ได้มีความยาวน้อยกว่า 7 หลักให้เติม 0 ต่อไปจนรหัสมีความยาวครบ 7 หลัก

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ส	ซ ศ ษ ส
ค	ค ฎ
ต	ด ฏ
ท	ฐ ท ฒ ถ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ ฤ ฦ
ว	ว
อ	อ
ฮ	ห ฮ

ตารางที่ 2.3 การกำหนดรหัสตัวอักษรของรหัสซาวด์เด็กซ์ภาษาไทย

รหัสตัวเลข	ตัวอักษร
0	ม ว ำ
1	ก บ ฃ ค ฅ ฆ
2	ง ย
3	ญ ณ น
4	ฎ ฏ ด ต ศ ษ ส
5	บ ป พ ภ
6	ผ ฝ ฟ ห อ ฮ
7	จ ฉ ช ซ ฌ
8	ฐ ฑ ฒ ถ ท ธ
9	ร ฤ ล ฦ ฬ

ตารางที่ 2.4 การกำหนดรหัสตัวเลขของรหัสชาวด์เด็กซ์ภาษาไทย

ส่วนข้อกำหนดเพื่อให้เหมาะสมกับภาษาไทยแบ่งดังนี้กรณีต่าง ๆ ดังนี้

- กรณีพบ ไ- ไ- ไ-ย และ -ัย จะเปลี่ยนให้อยู่ในรูปแบบเดียวกันคือ -ัย ก่อนทำการเข้ารหัสชาวด์เด็กซ์ เพื่อให้ได้รหัสชาวด์เด็กซ์เหมือนกัน เนื่องจากสระดังกล่าวอ่านออกเสียงเหมือนกัน เช่น ใท ไท ไทย และ ทัย จะเปลี่ยนเป็น ทัย
- กรณีพบ รร จะเปลี่ยนเป็น -ัน ในกรณีที่ไม่มีตัวสะกดตามหลัง และเปลี่ยนเป็น -ั กรณีที่มีตัวสะกดตามหลัง เช่น เปลี่ยนคำว่า สรรเพชร รังสรรค์ พรรณนา และ ธรรมรัตน์ เป็น สันเพชร รังสัน พัณนา และ ธัมรัตน์ ตามลำดับ
- กรณีพบการันต์ จะตัดการันต์และพยัญชนะที่มีตัวการันต์กำกับรวมทั้งสระและอักษรควบการันต์ทิ้ง เช่น คำว่า จันทร์ ศักดิ์ และ พันธุ์ เปลี่ยนเป็น จัน ศัก และ พัน ตามลำดับ

ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย ดังตารางที่ 2.5

ชื่อภาษาไทย	รหัสชาวด์เด็กซ์ภาษาไทย
อัมพร	อ059000
อำกรณ์	อ059000
พรรณศักดิ์	พ341000
พันธุ์ศักดิ์	พ341000
เนืองนิจ	น623400
เนืองนิตย์	น623400
สับปะรด	ส559400
สัปรส	ส594000

ตารางที่ 2.5 ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย

งานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา มีการออกแบบขั้นตอนวิธีการเข้ารหัสชาวด์เด็กซ์ภาษาไทยเพื่อให้รหัสที่ได้ตรงกันระหว่างคำที่สะกดถูกและสะกดผิด เพื่อแก้ปัญหาการสืบค้นคำไทยที่มักสะกดผิด โดยมีแนวคิดว่าพยายามพิจารณาอักขระทุกตัวในการเข้ารหัส หลักเกณฑ์ต่าง ๆ ในการเข้ารหัสชาวด์เด็กซ์มีดังนี้

- ใช้อักขระไทยทั้งหมดในการเข้ารหัส เพื่อสะดวกในการอ่าน
- ถ้าตัวอักษรที่อยู่ติดกันเป็นตัวควบกล้ำให้เปลี่ยนเป็นรหัสตัวเดียวกันคือเป็นตัวหน้าของตัวควบกล้ำ เช่น คำว่า ปรับ คลอง และ จริง ให้เข้ารหัสเฉพาะ ปับ คอง และจิง ตามลำดับ
- ถ้าตัวอักษรที่อยู่ติดกันเหมือนกันให้เปลี่ยนเป็นรหัสเพียงหนึ่งตัว
- ถ้าตัวอักษรที่อยู่ติดกันเป็นอักษรนำเสียงสนิทมี 9 ตัวได้แก่ อย หง หญ หน หม หย หร หล และ หว เช่นคำว่า อยู่ เหงือก หญิง หนู หมอ หยก หริ่ง หลวง และ หวัง เป็นต้น โดยจะตัดอักษร อ หรือ ห แล้วเปลี่ยนอักษรที่เหลือเป็นรหัสเพียงหนึ่งตัว
- ไม่นำวรรณยุกต์และไม้ไต่คู้มาเข้ารหัส
- เปลี่ยนพยัญชนะตัวแรกของคำเป็นรหัสตัวอักษรตามตารางที่ 2.6
- เปลี่ยนอักขระตั้งแต่ตัวที่สองขึ้นไปเป็นรหัสตามตารางที่ 2.7
- การเปลี่ยนตำแหน่งสระหน้า ได้แก่ เ- แ- โ- ไ- ใ- ไปไว้หลังสุดของคำ เพื่อให้คำที่สะกดผิดได้รหัสเหมือนกันเช่น ขโมย กับ โขมย ฉไล กับ ไฉล เป็นต้น

- ไม่นำอักขระ -ะ ั และ ิ มาเข้ารหัส
- สระ ุ ให้เข้ารหัสเป็น ู ยกเว้นสระ ุ อยู่ได้อักษร ด หรือ ธ ให้ตัดทิ้ง
- สระอื่นที่เหลือให้คงไว้

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ซ	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ฐ ท ฒ ถ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ฤ ฬ ฤๅ
ว	ว
อ	อ
ฮ	ห ฮ

ตารางที่ 2.6 การกำหนดรหัสอักษรสำหรับอักขระตัวแรกของรหัสซาวด์เด็กซ์ภาษาไทย

รหัสตัวอักษร	ตัวอักษร
ก	ก ข ค ฆ
ง	ง
ด	จ ฉ ช ซ ฌ ฎ ฏ ฐ ฑ ฒ ด ต ถ ท ธ ศ ส ษ
น	ญ ณ น ร ล ฬ
บ	บ ป พ ฟ ภ ผ ฝ
ม	ม ำ
ย	ย
ว	ว
ไ	ไ- ย -ย ใ-

ตารางที่ 2.7 การกำหนดรหัสอักษรสำหรับอักขระตัวถัดไปของรหัสซาวด์เด็กซ์ภาษาไทย

ตัวอย่างการเข้ารหัสซาวด์เด็กซ์ภาษาไทย ดังตารางที่ 2.8

ชื่อภาษาไทย	รหัสซาวด์เด็กซ์ภาษาไทย
พหูสูต	พหูดูด
พหุสุต	พหูดูด
กิตติพรรณ	กดบน
กิติพันธุ์	กดบน
เนืองนิจ	นืองนดเ
เนืองนิตย์	นืองนดเ
สัปปะรด	ซบนด
สัปรส	ซบนด
เถลไถล	ทนดนเไ
ถเลไถล	ทนดนเไ

ตารางที่ 2.8 ตัวอย่างการเข้ารหัสซาวด์เด็กซ์ภาษาไทย

งานวิจัยของ Theppitak Karoonboonyanan ได้นำเสนอวิธีใหม่ในการเข้ารหัสสำหรับซาวด์เด็กซ์ภาษาไทย โดยมีแนวคิดที่ว่าระบบหน่วยคำในคำในภาษาไทยไม่มีเครื่องหมายแบ่งพยางค์ ทำให้คำบางคำสามารถอ่านออกเสียงได้หลายรูปแบบ เช่น คำว่า "เสนา" สามารถอ่านออกเสียงได้ทั้ง เสนา และ เส-นา เป็นต้น ซึ่งการเข้ารหัสซาวด์เด็กซ์ได้ใช้หลักการอ่านออกเสียงของคำในการเข้ารหัส ถ้าหากว่าแบ่งพยางค์ผิดทำให้เสียงที่ได้ในการเข้ารหัสก็ผิดด้วย ดังนั้นเพื่อลดความผิดพลาดในการอ่านออกเสียงจึงได้ใช้เทคนิค ออโตมาตันสถานะจำกัดแบบไม่แน่นอน มีข้อมูลออก (Nondeterministic Finite Automaton with Output) เพื่อสร้างรหัสซาวด์เด็กซ์สำหรับการอ่านออกเสียงทุกกรณีของคำ ทำให้การเข้ารหัสคำหนึ่งคำอาจจะได้หลายรหัสซาวด์เด็กซ์ซึ่งจะต่างจากวิธีการเข้ารหัสซาวด์เด็กซ์แบบอื่นที่ได้กล่าวมาแล้วคือการเข้ารหัสคำหนึ่งคำจะได้รหัสซาวด์เด็กซ์เพียงรหัสเดียว ส่วนในการค้นคืนจะนำรหัสทุกตัวของคำที่เป็นข้อคำถามไปเปรียบเทียบกับรหัสทุกตัวของคำที่ต้องการค้นคืน ถ้าพบว่ามีรหัสเหมือนกันเพียงคู่หนึ่งก็จะถือว่าคำทั้งคู่เป็นคำที่อ่านออกเสียงคล้ายกัน การกำหนดรหัสซาวด์เด็กซ์ และขั้นตอนวิธีการเข้ารหัสซาวด์เด็กซ์มีดังนี้

การกำหนดรหัสซาวด์เด็กซ์สำหรับอักขระไทยมีดังนี้

1. กำหนดรหัสซาวด์เด็กซ์สำหรับพยัญชนะต้น จะแบ่งตามเสียงของพยัญชนะ 21 กลุ่มของไทย และได้รวม ร กับ ล เป็นกลุ่มเดียวกัน โดยจะใช้พยัญชนะไทยหนึ่งตัวแทนกลุ่มเสียง ดังแสดงในตารางที่ 2.9
2. กำหนดรหัสซาวด์เด็กซ์สำหรับสระ จะแบ่งตามเสียงของสระและรวมสระเสียงสั้นและยาวเข้าด้วยกัน โดยจะใช้อักขระโรมันกลุ่มเสียงสระดังแสดงในตารางที่ 2.10
3. กำหนดรหัสซาวด์เด็กซ์สำหรับตัวสะกด จะแบ่งกลุ่มเสียงตามมาตราต่าง ๆ ของอักษรไทย และได้เพิ่มกลุ่ม ฮ สำหรับคำที่ไม่มีตัวสะกดดังแสดงในตารางที่ 2.11

รหัสซาวด์เด็กซ์	ตัวอักษร
ก	ก
ค	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ซ	ซ ศ ษ ส
ด	ด ท* ฎ
ต	ต ฏ
ท	ฐ ท* ฒ ถ ท ธ
น	ณ น

รหัสซาวด์เด็กซ์	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ
ว	ว
อ	อ
ฮ	ห ฮ

* ท สามารถอ่านออกเสียงได้ 2 แบบคือ /ด/ และ /ท/ จึงสร้างรหัสทั้ง 2 แบบ

ตารางที่ 2.9 การเข้ารหัสสำหรับพยัญชนะต้น

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

รหัสชาวด์เด็กซ์	รูปแบบสระ
A	-ะ ั -รร- -อ ำ ใ- ไ- ไ-ย เ-า -า
I	ิ ี -ฤ- (ริ)
V	ึ ื -ฤ- (รึ)
U	ุ ู
E	เ-ะ เ็ เ-
X	แ-ะ แ็ แ-
O	โ-ะ อ+ตัวสะกด โ-
C	เ-าะ ็อ- -อ- -ร (อร)
D	เ-อะ เ-อ เ็ เ-ย
J	เ็ ยะ เ็ ย
W	เ็ อะ เ็ อ
S	ัวะ ัว -ว-

ตารางที่ 2.10 การเข้ารหัสสำหรับสระ

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

ตารางที่ 2.11 การเข้ารหัสสำหรับตัวสะกด

ขั้นตอนวิธีการเข้ารหัสชาวด์เด็กซ์มีดังนี้

1. สร้างโครงร่างเครื่องเข้ารหัสชาวด์เด็กซ์ จากการรวบรวมกฎเกณฑ์ต่าง ๆ ในการสร้างคำในภาษาไทย
2. ขยายเครื่องเข้ารหัสชาวด์เด็กซ์ โดยกำหนดข้อมูลนำออกให้สำหรับแต่ละกฎเกณฑ์
3. นำคำที่ต้องการเข้ารหัสผ่านเครื่องเข้ารหัสชาวด์เด็กซ์ ผลลัพธ์ที่สมบูรณ์แต่ละตัวที่ได้คือรหัสชาวด์เด็กซ์

ตัวอย่างขั้นตอนการสร้างเครื่องเข้ารหัสชาวด์เด็กซ์

สมมุติว่ากฎเกณฑ์การสร้างคำที่ถูกต้องมีรูปแบบดังนี้

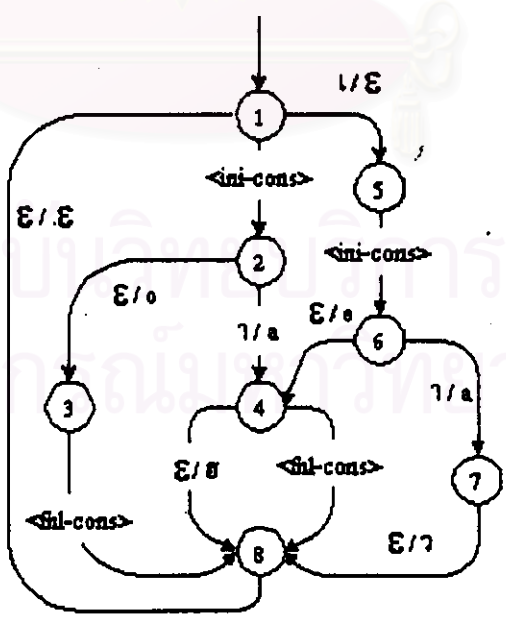
1. <int-cons> <fnl-cons>
2. <int-cons> า [<fnl-cons>]
3. เ <int-cons> า
4. เ <int-cons> [<fnl-cons>]

โดยที่ <ini-cons> ::= <cons> | <cons> <cons>

<cons> ::= ส | น

<fnl-cons> ::= ส | น

จากกฎเกณฑ์ดังกล่าวสามารถแสดงดังในรูปที่ 2.2



รูปที่ 2.2 โครงร่างเครื่องเข้ารหัสชาวด์เด็กซ์

กำหนดข้อมูลนำออกให้แต่ละกฎเกณฑ์ดังนี้

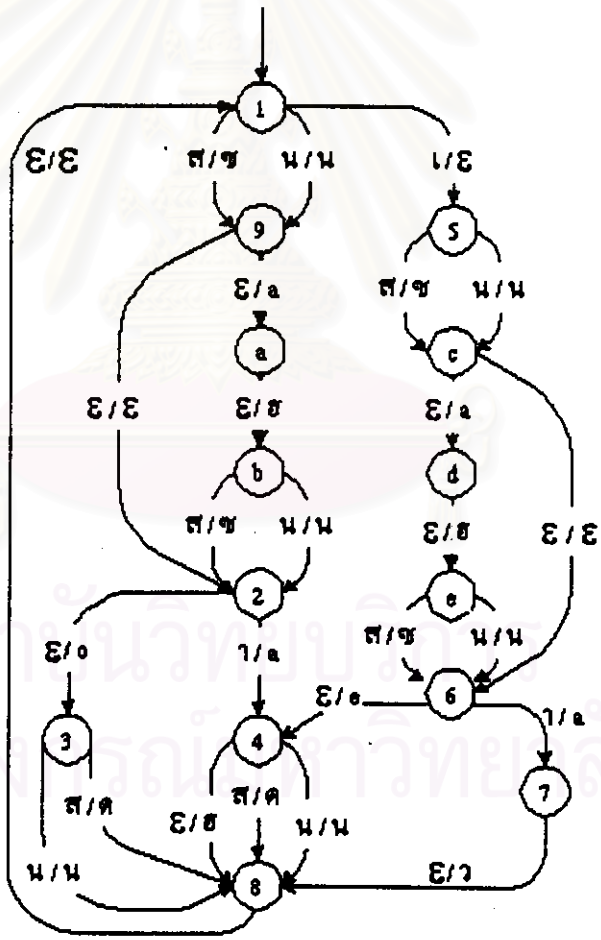
1. <int-cons> (ε / o) <fnl-cons>
2. <int-cons> (า / a) (ε / ย)
3. <int-cons> (า / a) <fnl-cons>
4. (เ / ε) <int-cons> (า / a) (ε / ว)
5. (เ / ε) <int-cons> (ε / e) (ε / ย)
6. (เ / ε) <int-cons> (ε / e) <fnl-cons>

โดยที่ <ini-cons> ::= <cons> | <cons> (ε / a) (ε / ย) <cons>

<cons> ::= (ส / ช) | (น / น)

<fnl-cons> ::= (ส / ค) | (น / น)

จากการกำหนดข้อมูลนำออกสามารถแสดงดังในรูปที่ 2.3



รูปที่ 2.3 เครื่องเข้ารหัสชาวด์เด็กซ์

ตัวอย่างนำคำว่า "เสนา" ผ่านเครื่องเข้ารหัสชาวด์เด็กซ์ ได้ผลลัพธ์ที่สมบูณณ์สองค่าคือ

1. (1)—เ / ɛ →(5)—ส / ช →(c)—ɛ / a →(d)—ɛ / อ →(e)—น / น →(6)→ า / a →(7)—ɛ / ว →(8)*
2. (1)—เ / ɛ →(5)—ส / ช →(c)—ɛ / ɛ →(6)—ɛ / e →(4)—ɛ / อ →(8)→ ɛ / ɛ →(1)→ น / น →(9)—ɛ / ɛ →(2)—า / a →(4)—ɛ / อ →(8)*

สรุปรหัสชาวด์เด็กซ์ที่ได้คือ "ชaอนaว" และ "ชeนอaอ"

ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย ดังตารางที่ 2.12

ชื่อภาษาไทย	รหัสชาวด์เด็กซ์ภาษาไทย
บุญญา	บนนยaอ, บนอยoนยaอ, บนนยaอยaอ, บนอยcอยaอ, บนอยaอยaอ, บนนยaอ
บุณยา	บนนยaอ, บนนนaอยaอ, บนอนcอยaอ, บนอนaอยaอ
สับปะรด	ชaบปaอฉoด, ชaบบaอปaอถoคคaอ, ชaบปaอถoคคaอ, ชaบบaอปaอถcอคcอ,ชaบบaอปaอถaอคcอ, ชaบปaอถcอคcอ, ชaบปaอถaอคcอ, ชaบบaอถcอคaอ, ชaบบaอปaอถaอคaอ, ชaบปaอถcอคaอ, ชaบปaอถaอคaอ, ชaบบaอปaอถoค
สัปรส	ชaบปaอฉoด, ชaบปaอถoคชaอ, ชaบถoคชaอ, ชaบปaอชcอ, ชaบชcอ, ชaบปaอถcอชcอ, ชaบปaอถaอชcอ, ชaบถcอชcอ, ชaบถaอชcอ, ชaบปaอชaอ, ชaบชaอ, ชaบปaอถcอชaอ, ชaบปaอถaอชaอ, ชaบถcอชaอ, ชaบถaอชaอ, ชaบถoค
ธรรมะ	ทaมมaอ, ทaนมaอ, ทcนถoมมaอ, ทcนถcอมaอ, ทcนถaอมaอ, ทcอถcอถoมมaอ, ทaอถcอถoมมaอ, ทcอถcอถoมมaอ, ทaอถaอถoมมaอ, ทcอถcอถcอมaอ, ทaอถcอถcอมaอ, ทcอถaอถcอมaอ, ทcอถcอถaอมaอ, ทaอถaอถcอมaอ, ทaอถcอถaอมaอ, ทcอถaอถaอมaอ, ทaอถaอถaVมaอ
ธัมมะ	ทaมมaอ, ทaมมaอมaอ

ตารางที่ 2.12 ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย

2.7 ขั้นตอนวิธีระยะแก้ไขสั้นที่สุด (Minimum Edit Distance)

ระยะแก้ไขสั้นที่สุด [15] เป็นเทคนิคหนึ่งในการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระ ซึ่งจะทำการคำนวณหาจำนวนคำสั่งน้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้สายอักขระทั้งสองสายเหมือนกัน ตัวอย่างเช่น ระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE เท่ากับ 3 ซึ่งมีวิธีการคำนวณดังนี้

1. การลบตัวอักษร S	EXSAMBL	=>	EXAMBL
2. การแทนที่ตัวอักษร B ด้วย P	EXAMBL	=>	EXAMPL
3. การเพิ่มตัวอักษร E	EXAMPL	=>	EXAMPLE

จากวิธีการคำนวณข้างต้นสามารถเขียนในอยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ได้ดังนี้

$$\text{Edit}(P_0, W_0) = 0$$

$$\text{Edit}(P_j, W_0) = j$$

$$\text{Edit}(P_0, W_k) = k$$

$$\text{Edit}(P_j, W_k) = \min[\ \text{Edit}(P_{j-1}, W_k) + 1,\ \text{Edit}(P_j, W_{k-1}) + 1,\ \text{Edit}(P_{j-1}, W_{k-1}) + r(p_j, w_k)\ ]$$

โดยที่

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

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

$r(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

1 ถ้า p_j ไม่เท่ากับ w_k

---

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

ขั้นตอนวิธีการคำนวณหาค่าระยะการแก้ไขสั้นที่สุดใช้เวลาทำงานเป็น O(mn) โดย m และ n คือความยาวของสายอักขระที่ 1 และ 2 ตามลำดับ เมื่อใช้เทคนิคกำหนดการพลวัต (Dynamic Programming) [16] ดังแสดงในรูปที่ 2.4

```
int EitDist(char *P; char *W) {
  int  F [STRLEN+1][STRLEN+1];
  int   i, j, l1, l2;

  l1 = strlen (P);
  l2 = strlen (W);
  for (i=0; i <= STRLEN; i++)
    F[0][i] = F[i][0] = i;
  for (i=0; i <= STRLEN; i++)
     for (j=0; j <= STRLEN; j++)
        F[i][j] := Min(   F[i-1][j]    +  1,
                          F[i1][j-1]   +  1,
                          F[i-1][j-1]  + Equal( P[i-1], W[j-1] ) );
  return (F[l1][l2]);
}

int  Equal(char a, char b) {
   return (  (a == b) ? 0 : 1 );
}
```

รูปที่ 2.4 โปรแกรมระยะแก้ไขสั้นที่สุด โดยใช้เทคนิคกำหนดการพลวัต

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

2.8 สรุป

ในบทนี้ได้กล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้องในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ โดยใช้แนวคิดในการแปลงข้อคำถามที่เป็นคำทับศัพท์ด้วยทฤษฎีของการถอดอักษร และใช้เทคนิคซาวด์เด็กซ์ช่วยในการเข้ารหัสคำ โดยรายละเอียดต่าง ๆ จะกล่าวไว้ในบทถัดไป

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย