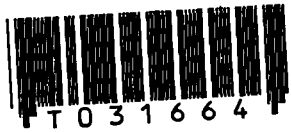
การตัดคำไทยโดยใช้ดิกชันนารีที่มีโครงสร้างข้อมูลแบบแฮชชิ่ง

THAI WORD SEPARATOR USING HASHING DICTIONARY



สรศักดิ์ ไทยแท้

SORASAK THAITAE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2541

ISBN 974 - 622 - 303 - 8

เลขหมู่.....
เลขทะเบียน 31664
วัน, เดือน, ปี 21 ธ.ค. 2541

THAI WORD SEPARATOR USING HASHING DICTIONARY

SORASAK THAITAE

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
1998

ISBN 974 - 622 - 303 - 8

COPYRIGHT 1998
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

หัวข้อวิทยานิพนธ์	การตัดคำไทยโดยใช้ดิกชันนารีที่มีโครงสร้างข้อมูลแบบแฮชชิ่ง
นักศึกษา	นายสรศักดิ์ ไทยแท้
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ.ดร.บุญธีร์ เครือตราชู
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
พ.ศ.	2541

บทคัดย่อ

เนื่องจากตัวอักษรไทยมักจะเขียนติดกันไม่มีตัวอักษรพิเศษขั้นระหว่างคำดังในภาษาอังกฤษ จึงเป็นการยากที่จะทราบว่าผู้เขียนต้องการจะแบ่งคำอย่างไร หรือ คำใดมีคำที่สะกดผิดอยู่ ทั้งนี้การรู้ขอบเขตของคำโดยอัตโนมัติ จะช่วยให้สามารถตรวจสอบคำสะกดได้ถูกต้องและรวดเร็วขึ้นทั้งสามารถทำ index จากคำในข้อความได้ งานวิจัยนี้ได้นำเสนออัลกอริทึมในการสืบค้นคำศัพท์และพยายามตัดแบ่งคำโดยอัตโนมัติโดยใช้พจนานุกรม (Dictionary) ที่มีโครงสร้างแบบแฮชชิ่ง (Hashing) เป็นตัวช่วยในการตัดคำ รวมทั้งได้ออกแบบอัลกอริทึมแบบ Iterative และโครงสร้างข้อมูล Stack ที่ใช้ในการเก็บตำแหน่งคำในการตัดคำแทนแบบเดิมที่เป็น Recursive รวมทั้งพจนานุกรมที่ใช้แบบ Hashing ทำให้สามารถเพิ่มเติมคำได้ง่ายรวมทั้งสามารถเลือกหาคำคล้ายได้สะดวก งานวิจัยนี้ยังได้ทำการเปรียบเทียบการตัดคำและการตรวจสอบคำผิดกับโปรแกรม Commercial คือ Word 6.0 และ Word 7.0 ของบริษัทไมโครซอฟต์รวมทั้งได้รวบรวม ข้อผิดพลาดของ Word 6.0 และ Word 7.0 มาทำการปรับปรุงอัลกอริทึมในงานวิจัยนี้ งานวิจัยนี้มีข้อดีคือสามารถตัดคำได้ทุกตำแหน่ง ถ้าคำๆนั้นมีในพจนานุกรมซึ่งจะพบคำในประโยคที่ถูกต้องตามพจนานุกรมและยังใช้ความยาวมากที่สุดของคำไทยในพจนานุกรม (25 ตัวอักษร) มาช่วยในการเพิ่มความเร็วในการตัดคำโดยทำการพิจารณาคำทีละ 25 ตัวอักษร นอกจากนี้ยังเพิ่มเติมโครงสร้างพจนานุกรมนี้ให้มีการหยุดตัดคำที่สั้นกว่า 25 ตัวอักษรเมื่อพบว่าคำๆนั้นไม่มีตัวอักษรต่อในพจนานุกรม แต่มีข้อเสียคือในการค้นหาคำจะช้ากว่าโครงสร้างข้อมูลแบบทรัย (Trie)

Thesis Title	Thai Word Separator Using Hashing Dictionary
Student	Mr.Sorasak Thaitae
Thesis Advisor	Assistant Professor Dr.Boontee Kruatrachue
Level of Study	Master of Science program in Information Technology
Degree	Faculty of Information Technology
	King Mongkut's Institute of Technology Ladkrabang
Year	1998

ABSTRACT

Thai alphabet are written adjacent, without any special alphabet between words. Therefore, it is difficult to separate each word in a line, or to detect any misspelling word. Separate words automatically will help both spelling correction and also document indexing. This research present search algorithm and the new Thai word separater using Hashing dictionary, and design the new separator algorithm using iterative instead of recursive with double stack data structure. We can also make word suggestion to the incorrect word with little modification to the dictionary. Hashing dictionary is also easier to add new words. The new algorithm and dictionary has been compared with softwares Word 6.0 and Word 7.0 of Microsoft in both the running time and correctness. The word separable result is correct as long as the words are in the dictionary. The separable speed has been improved by noticing the max number of character per word in the dictionary is 25. The dictionary also mark ended words so that the separator can stop before 25 characters when the ended words are detected. The speed can be improve with trie structure with trade of new words addition flexibility and word suggestion feature.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยดีเพราะได้รับความเมตตาจาก ผศ. ดร. บุญธีร์ เครือตราชู ที่ได้ให้ความกรุณาสละเวลา เพื่อให้คำแนะนำและปรึกษาแนวทางการแก้ปัญหาต่าง ๆ ตลอดการทำวิจัยนี้ ผู้วิจัย ขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณ บิดา, มารดา, ญาติ ๆ และ คุณเสริมศรี สุกรินทร์ ที่ให้กำลังใจในการทำวิจัย

ขอขอบคุณ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ได้เอื้อเฟื้อรายการคำศัพท์ภาษาไทยฉบับราชบัณฑิตยสถาน พ.ศ. 2525

ขอขอบคุณ คุณฉันทนา อุบลพรรณ บรรณารักษ์ห้องสมุด มหาวิทยาลัยมหิดล ที่ได้จัดเตรียมข้อมูลต่าง ๆ ในงานวิจัย

ขอขอบคุณ คุณภวาลัย ไกรพีรพรรณ, สิทธิศักดิ์ ทองสุข, วรวิทย์ พูลสวัสดิ์ ที่กรุณาให้คำแนะนำเรื่องการเขียนโปรแกรมภาษา C เพื่อเป็นโปรแกรมที่ใช้ในการทดสอบของงานวิจัยนี้

ขอขอบคุณ คุณเกษรา เวียงวะลัย และคุณอนุรัตน์ บูรณะเรศน์ ที่ได้เอื้อเฟื้ออุปกรณ์ในการทำวิจัย

ขอขอบคุณ เพื่อน ๆ วิทยาการคอมพิวเตอร์ รุ่น 1 และพี่ ๆ ทุกคนในคณะเทคโนโลยีสารสนเทศ ที่ให้กำลังใจและช่วยเหลือในงานวิจัยนี้

ขอขอบคุณพี่บุญช่วย และ พี่ ๆ เจ้าหน้าที่คณะเทคโนโลยีสารสนเทศที่ให้การช่วยเหลือด้านต่าง ๆ

สรศักดิ์ ไทยแท้

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... I
บทคัดย่อภาษาอังกฤษ..... ...II
กิตติกรรมประกาศ..... III
สารบัญ..... IV
สารบัญตาราง.....VII
สารบัญภาพ... VIII
บทที่
1. บทนำ..... ..1
ความเป็นมาของงานวิจัย.....1
ผลงานที่ผ่านมา.....1
แนวทางของวิทยานิพนธ์..... ...6
วัตถุประสงค์ของงานวิจัย..... ...7
โครงสร้างของวิทยานิพนธ์..... ...7
2. การค้นหาคำในพจนานุกรมตามโครงสร้างข้อมูลแบบต่างๆ.....8
โครงสร้างข้อมูลพจนานุกรมอิเล็กทรอนิกส์..... ...8
โครงสร้างข้อมูลแบบแฮชชิ่ง.....12
3. แนวทางการออกแบบพจนานุกรมอิเล็กทรอนิกส์..... .18
ประเภทของพจนานุกรมอิเล็กทรอนิกส์..... .18
ลักษณะคำไทย.....20
โครงสร้างของพยางค์ในภาษาไทย..... .20
ชนิดของหน่วยคำในภาษาไทย.....21
การวิเคราะห์ตัวอักษรในภาษาไทย..... .22
การออกแบบพจนานุกรมอิเล็กทรอนิกส์.....24

สารบัญ (ต่อ)

หน้า

บทที่

โครงสร้างในการเก็บคำในพจนานุกรมอิเล็กทรอนิกส์ 25
ขั้นตอนการออกแบบและการสร้างพจนานุกรมอิเล็กทรอนิกส์ 25
ผลของการเก็บโครงสร้างพจนานุกรมแบบแฮชชิ่ง....29
อัลกอริทึมในการค้นหาคำ....37
การออกแบบโครงสร้างพจนานุกรมในการค้นหาคำคล้าย....39
อัลกอริทึมในการค้นหาคำคล้าย....41
4. การตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์....43
การวิเคราะห์โครงสร้างพยางค์คำของภาษาไทยในการตัดคำ....44
กฎการตัดคำโดยพิจารณาจากอักษรวิธี....45
การวิเคราะห์การตัดแบ่งคำโดยการตรวจสอบจากพจนานุกรม....46
ผลงานวิจัยการตัดแบ่งโดยการตรวจสอบจากพจนานุกรม....47
หลักการตัดแบ่งคำภาษาไทยโดยใช้โครงสร้างพจนานุกรมแบบแฮชชิ่ง....49
การถอยหลังเมื่อพบปัญหาหรือคำที่คลุมเครือ (BackTracking)....52
อัลกอริทึมในการตัดแบ่งคำภาษาไทย....54
การตัดแบ่งคำภาษาไทยเมื่อพบคำผิดในเอกสาร....56
5. ผลการทดสอบการตัดแบ่งคำและเปรียบเทียบกับ Microsoft Word....57
ข้อมูลที่ใช้ในการทดสอบ....57
ผลการทดสอบการตัดคำโดยใช้พจนานุกรมกับ Microsoft Word....58
6. บทสรุปและข้อเสนอแนะ....65
บทสรุป....65
ข้อเสนอแนะ....66
บรรณานุกรม....67

สารบัญ (ต่อ)

หน้า

ภาคผนวก 70
ภาคผนวก ก ตารางแอสกี (ASCII) 72
ภาคผนวก ข ตารางรายการคำที่ MICROSOFT WORD 6.0
ไม่สามารถตัดแบ่งคำได้ 74
ตารางรายการคำที่ MICROSOFT WORD 7.0
ไม่สามารถตัดแบ่งคำได้ 94
ภาคผนวก ค บทความทางวิชาการ 111
ประวัติผู้เขียน 121

สารบัญตาราง

ตารางที่ หน้า

3.1 แสดงการเกิด Collision.....30
3.2 แสดงขนาดตัวอักษรที่ประกอบเป็นคำ..... ..31
3.3 แสดงการแจกแจงการใช้ตัวอักษรเป็นตัวแรก.....32
3.4 แสดงการแจกแจงการใช้ตัวอักษรเป็นตัวสุดท้าย..... ..33
3.5 แสดงการแจกแจงความถี่ของตัวอักษร..... ..34
3.6 แสดงการแจกแจงความถี่ของตัวอักษรที่ใช้มากตามลำดับ.....35
3.7 แสดงจำนวนคำที่ตัวอักษรประกอบเป็นคำมากที่สุด 10 อันดับ..... ..37
5.1 แสดงจำนวนคำที่สามารถตัดแบ่งได้และไม่สามารถตัดแบ่งได้..... ..58
5.2 แสดงความเร็วในการตัดคำ..... ..59

สารบัญภาพ

ภาพที่ หน้า

1.1 แสดงขั้นตอนการตัดแบ่งคำและการตรวจสอบตัวสะกด..... 2
1.2 แสดงโครงสร้างพจนานุกรมแบบ Table-Table-Index-Search.....3
1.3 แสดงโครงสร้างพจนานุกรมแบบ Table-Index-Search..... 4
1.4 แสดงโครงสร้างพจนานุกรมแบบ Tree-Structure.....4
1.5 แสดงการเก็บคำแบบ Common Factor..... 5
1.6 แสดงโครงสร้างต้นไม้ในการเก็บคำแบบ Common Factor..... 5
2.1 แสดงโครงสร้างข้อมูลแบบทรัย Trie Structure..... .. 11
2.2 แสดงการแก้ปัญหาการชนกันโดยวิธี Open Addressing..... .. 16
2.3 แสดงการแก้ปัญหาการชนกันโดยวิธี Chaining..... 17
3.1 แสดงโครงสร้างการเก็บคำของพจนานุกรม..... .. 25
3.2 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง..... .. 26
3.3 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง..... .. 28
3.4 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง..... .. 29
3.5 แสดงวิธีการสร้างพจนานุกรมคำคล้าย..... .. 40
3.6 แสดงวิธีการค้นหาคำคล้ายในพจนานุกรม..... .. 41
4.1 แสดงค่าตำแหน่งที่สามารถทำการตัดได้ใน Stack..... .. 50
4.2 แสดงค่าตำแหน่งที่สามารถทำการตัดได้ใน Stack 51
4.3 แสดงค่าตำแหน่งที่สามารถทำการตัดได้ใน Stack 52
4.4 แสดงขั้นตอนตำแหน่งการตัดคำใน Stack ในการทำ BackTracking..... .. 53

บทที่ 1

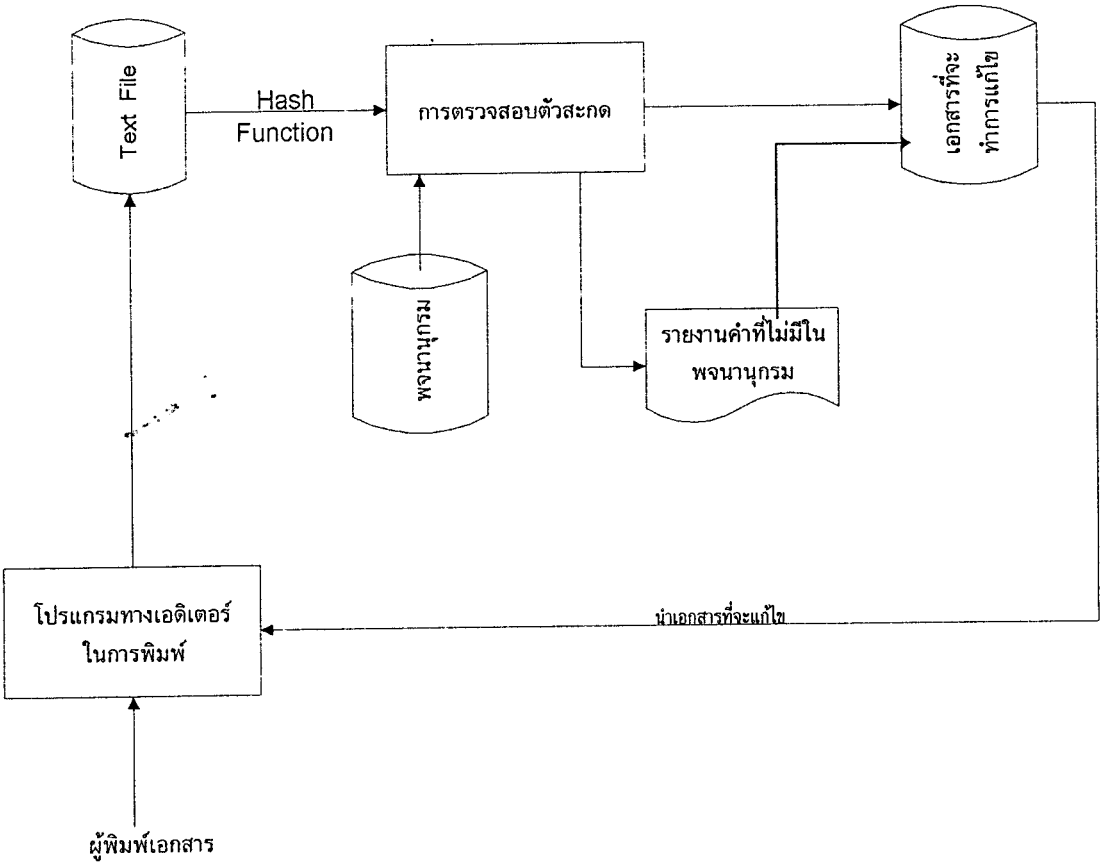
บทนำ

ความเป็นมาของงานวิจัย

ในการพัฒนาการทางด้านคอมพิวเตอร์ เพื่อใช้ในการประมวลผลข้อมูล ได้มีความต้องการมากขึ้น ซึ่งในปัจจุบัน ได้มีการนำเอาคอมพิวเตอร์มาใช้ในการประมวลผลทางด้านภาษามากขึ้น โปรแกรมการตราจสอบตัวสะกดและการตัดแบ่งพยางค์ ถ้าในภาษาไทยก็เป็นส่วนหนึ่งของการประมวลผลทางด้านภาษา โดยประโยคในภาษาไทยนั้น จะมีการเขียนคำที่ต่อเนื่องติดกัน โดยไม่มีตัวอักษรพิเศษขึ้นระหว่างคำ จนกว่าจะจบประโยค หรือมีการเว้นวรรค หรือ ขึ้นย่อหน้าใหม่ ซึ่งมีความแตกต่างจากคำในประโยคภาษาอังกฤษ ที่มีการแบ่งแยกคำที่ชัดเจน ทำให้เป็นการยากที่จะทำให้ทราบว่าเราควรจะตัดแบ่งคำอย่างไร หรือ มีคำใดที่มีคำสะกดผิดอยู่ในประโยค การทราบขอบเขตของคำในประโยคจึงมีความจำเป็นอย่างยิ่งในการตัดแบ่งคำดังตัวอย่างประโยค "เธอมารอกลับบ้าน" เราสามารถทำการตัดแบ่งได้ดังนี้ เธอ * มา *รอ * กลับ * บ้าน หรือ เธอ * มาร * อก * ลับ * บ้าน หรือ เธอ * มา * รอก * ลับ * บ้าน ซึ่งสามารถแบ่งได้หลายกรณี โดยการตรวจสอบและการตัดแบ่งคำที่มีขั้นตอน ดังภาพที่ 1.1

ผลงานวิจัยที่ผ่านมา

ผลงานวิจัยในด้านการตัดแบ่งคำภาษาไทยและการสร้างพจนานุกรม อิเล็กทรอนิกส์ ได้มีนักวิจัยผู้มีคุณวุฒิหลายท่านทำการศึกษาค้นคว้าไว้มากมาย ซึ่งจะขอยกตัวอย่างและสรุปวิธีการทำงานพอสังเขป ดังนี้



ภาพที่ 1.1 แสดงขั้นตอนการตัดแบ่งคำ และการตรวจสอบตัวสะกด

1. ผลงานในการสร้างพจนานุกรมภาษาไทย

1.1. ผลงาน วิจัยของ อาจารย์ ยืน ภู่วรวรรณ และวิวรรธน์ อิ่มอารมณ์ เรื่อง "การตรวจสอบตัวสะกดด้วยคอมพิวเตอร์" [9] ซึ่งได้นำเสนอโครงสร้างพจนานุกรมไว้ 3 แบบ คือ

1.1.1. Table - Table - Linear - Search

โครงสร้างวิธีนี้จะเก็บข้อมูลแบบเรียงต่อกัน

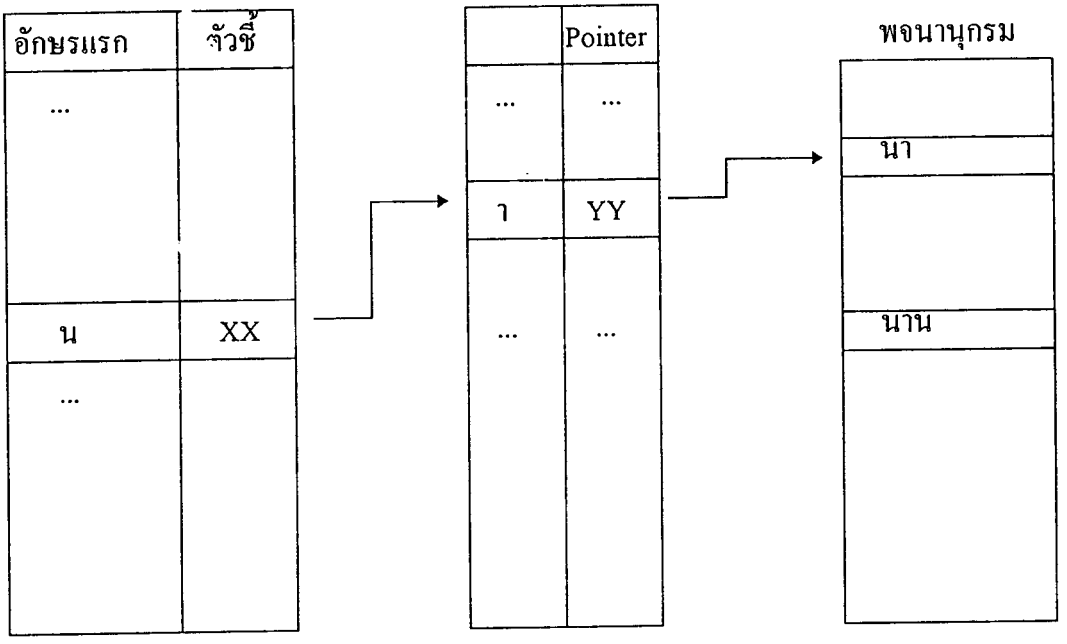
ก _____ 2กา 3กาก 3กาง 6กางเกง 3กาจ 5กาชาด _____

2กา หมายถึง เลขที่อยู่หน้าคำจะบอกจำนวนตัวอักษรที่ประกอบเป็นคำ

ด้วยวิธีการเรียงพจนานุกรมใช้วิธี Table - Table จะใช้ตารางนำหน้า 2 ตัวอักษร ดังนั้นโครงสร้างในพจนานุกรมจะลด 2 ตัวอักษรแรก ดังภาพที่ 1.2

ก _____ 0 1ก 4งเกง 1จ 3ชาด _____

ถ้าเป็นเลข 0 จะหมายถึงคำ 2 ตัวแรกเป็นคำในพจนานุกรม ตัวอื่นจะบอกความหมายของความยาวของคำในพจนานุกรมไปในตัว โดยลบออกจากสองตัวแรก



ตารางระดับ 1
ก-ย เ แ โ ไ ใ

ตารางระดับ 2
จำนวน 49 ตารางๆละ 68 ตัว

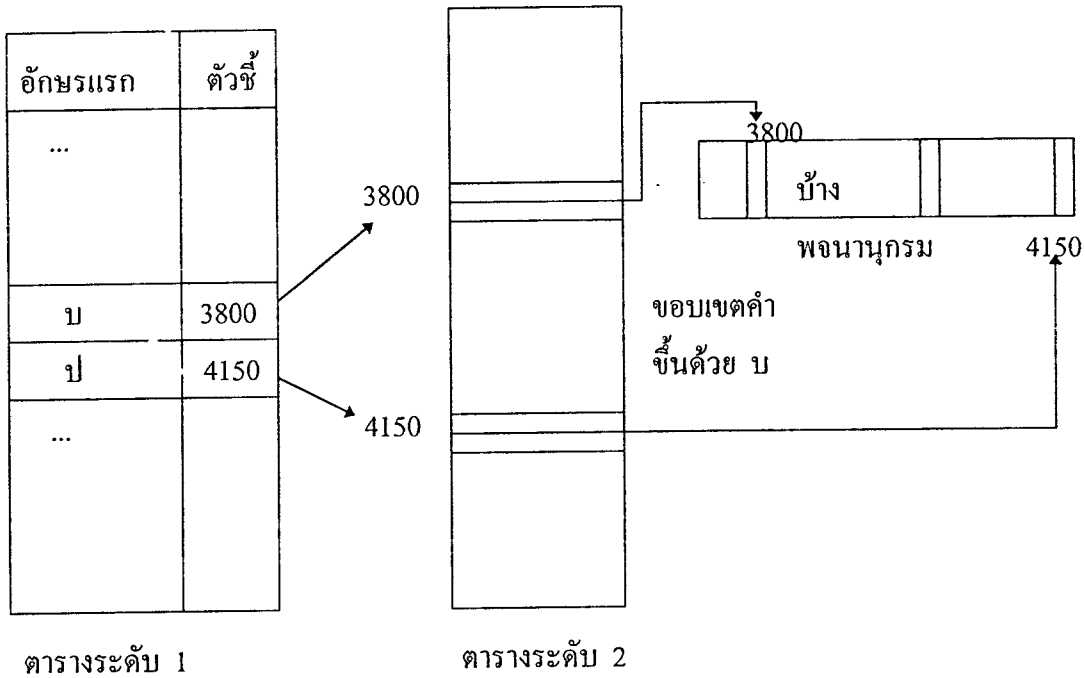
ภาพที่ 1.2 แสดงโครงสร้าง พจนานุกรม แบบ Table - Table - Index - Search

1.1.2. Table - Index - Search

โครงสร้างข้อมูลแบบนี้ ยึดหลักการให้ง่ายต่อการลดขนาด โดยเก็บข้อมูลดิกชันนารีมาเรียงติดต่อกัน โดยมีตัวเลขเป็นตัวคั่นระหว่างคำ ดังนี้

_____ 2กา 3กาก 3กาง 6กางเกง 3กาจ 5กาชาด _____

โดยจะเก็บโครงสร้าง ดังภาพที่ 1.3

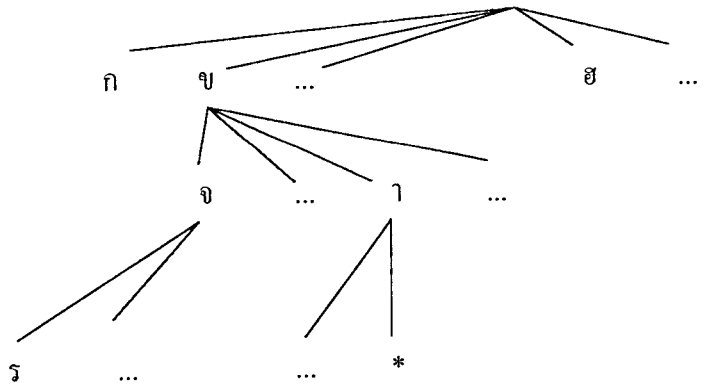


ภาพที่ 1.3 แสดงโครงสร้างการเก็บพจนานุกรมแบบ Table - Index - Search

ในการค้นหา เช่น ต้องการคำว่า บ้าง ซึ่งขึ้นต้นด้วยตัว บ การมองตารางระดับ 1 จะบอกได้ว่ากลุ่มอักษรขึ้นต้นด้วยตัว บ อยู่ที่ตำแหน่ง 3800-4150 ดังนั้น เราจะทำ binary search ในตัวชี้ในตารางระดับ 2 ตั้งแต่ 3800-4150 เพื่อหาคำในพจนานุกรมที่เก็บไว้

1.1.3. Tree Structure

โครงสร้างการเก็บข้อมูลเป็นแบบต้นไม้ (Tree) ดังภาพที่ 1.4



ภาพที่ 1.4 แสดงโครงสร้างการเก็บพจนานุกรม แบบ Tree Structure

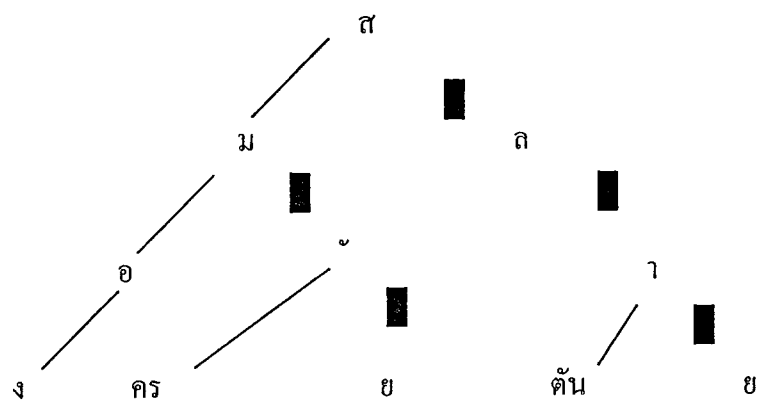
1.2. ผลงานวิจัยของ ผศ. ทวีศักดิ์ จันทร์วิทยานุชิต เรื่อง "การตรวจคำผิดในเอกสารภาษาไทย" [2] โดยงานวิจัยนี้เก็บโครงสร้างพจนานุกรมแบบต้นไม้ (Tree) โดยมีการนำเสนอวิธีการที่ใช้เลียนแบบ Common Factor ออกมานอกวงเล็บ อย่างพีชคณิต มาทำการจัดเรียงคำในพจนานุกรม ดังภาพที่ 1.5 และ 1.6 คำในพจนานุกรม คือ

สม สมอ สมอง สมัคร สมัย สลาตัน สลาย

Common Factor

1. ส(ม มอ มอง มัคร มัย ลาตัน ลาย)
2. ส(ม($ อ อ ั คร ั ย) ล(าตัน าย)
3. ส(ม($ อ($ ง) ั (คร ย)) ล(า(ตัน ย)))

ภาพที่ 1.5 แสดงการเก็บคำแบบ Common Factor



ภาพที่ 1.6 แสดงโครงสร้างต้นไม้ในการเก็บแบบ Common Factor

2. ผลงานการตัดแบ่งคำภาษาไทย

2.1. ผลงานวิจัย ของ นายสุรินทร์ จรรยาพรพงษ์ เรื่อง " A Thai Syallabe Separation Algorithm " งานวิจัยนี้ได้นำกฎทางอักขรวิธีมาทำการตัดแบ่งคำภาษาไทย โดยมีการกำหนดกฎเกณฑ์ในการตัดแบ่ง 62 ข้อ[22]

2.2. ผลงานวิจัย ของ อาจารย์ ยืน ภู่วรวรรณ, สมนึก คีรีโต และสุรศักดิ์ สงวนพงษ์ เรื่อง " การตัดแบ่งพยางค์คำไทยด้วยโครงสร้างข้อมูลแบบต้นไม้ " ผลงานวิจัยได้นำการพัฒนาปรับปรุงผลงานวิจัย ของ นายสุรินทร์ จรรยาพรพงษ์ มาทำการเก็บกฎการตัดแบ่งคำทั้ง 62 ข้อ เป็นโครงสร้างข้อมูลแบบต้นไม้ [10]

2.3. ผลงานวิจัย ของ อาจารย์ ยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ เรื่อง " การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี " เป็นงานวิจัยที่นำเสนอโดยเก็บคำในพจนานุกรม มาทำการเปรียบเทียบกับพจนานุกรม และได้นำเสนอการตัดแบ่งคำที่กำกวม โดยที่มีการย้อนกลับ (Backtrack) ไปตำแหน่งก่อนหน้าเพื่อหาคำที่จะนำมาเปรียบเทียบกับพจนานุกรมต่อ[8]

2.4. ผลงานวิจัย ของ นายสมนึก คีรีโต เรื่อง " ระบบฐานความรอบรู้ เพื่อการวิเคราะห์การแบ่งแยกพยางค์ไทย " [13] โดยนำเสนอการพัฒนาโปรแกรมด้วยภาษา Prolog โดยยึดหลักการตัดแบ่งคำที่ใช้กฎของ นายสุรินทร์ จรรยาพรพงษ์

2.5. ผลงานวิจัย ของ อาจารย์ ยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ เรื่อง " การตรวจสอบตัวสะกดภาษาไทยด้วยคอมพิวเตอร์ " นำเสนอโดยใช้โครงสร้างข้อมูลที่ท่านอาจารย์ได้ทำการวิจัยมาก่อนหน้า โดยการเก็บข้อมูลคำในพจนานุกรมแบบ Table - Table - Linear - Search [9]

2.6. ผลงานวิจัย ของ ดร. รัตติกร วรากูลศิริพันธ์และสมศักดิ์ จันวัน เรื่อง " การตัดคำจากประโยคในภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด " โดยนำเสนอการเก็บคำในพจนานุกรมเป็นแบบต้นไม้ (Tree) และนำเสนอการตัดแบ่งคำโดยการคำนวณ Weight ของประโยคที่จะนำมาตัดแบ่ง แล้วทำการเปรียบเทียบกับข้อมูลในพจนานุกรม [12]

<u>แนวทางของวิทยานิพนธ์</u>

จากผลงานวิจัยการตัดแบ่งคำในภาษาไทย การตรวจสอบตัวสะกด และการสร้างพจนานุกรม อิเล็กทรอนิกส์ ที่ผ่านมา มี 2 แนวทาง คือ วิธีการตรวจสอบตัวสะกด โดยใช้หลักการทางอักขรวิธีและวิธีการตรวจสอบตัวสะกดโดยใช้หลักการเปรียบเทียบกับพจนานุกรม เนื่องจากวิธีการตรวจสอบทางอักขรวิธีนี้ จะสามารถตรวจสอบได้รวดเร็ว แต่ความถูกต้องจะน้อย

เพราะอาจจะพบคำบางคำที่เป็นไปตามกฎแต่ไม่มีความหมาย ซึ่งจะพบค่อนข้างมาก ส่วนวิธีการตรวจสอบโดยการเปรียบเทียบกับพจนานุกรมจะมีความถูกต้องสูง แต่ความเร็วจะช้า ส่วนการสร้างพจนานุกรมอิเล็กทรอนิกส์ที่มีการวิจัยที่ผ่านมา จะเป็นการเก็บข้อมูลคำในพจนานุกรม เป็นแบบโครงสร้างข้อมูลแบบต้นไม้ (Tree) ซึ่งจะมีการปรับปรุงโครงสร้างยาก เมื่อมีการเพิ่มคำในพจนานุกรม

วิทยานิพนธ์ฉบับนี้ขอนำเสนอการสร้างโครงสร้างการเก็บคำในพจนานุกรมที่เหมาะแก่การตัดคำ โดยมีการเก็บคำแบบแฮชชิ่ง (Hashing Dictionary) ซึ่งโครงสร้างนี้จะทำให้ค้นหาคำได้รวดเร็ว และยังสะดวกต่อการปรับปรุงพจนานุกรม เมื่อเกิดคำใหม่ และยังได้นำเสนอวิธีการตัดแบ่งคำโดยใช้อัลกอริทึมแบบ Hashing มาช่วยในการตัดแบ่งคำและขอบเขตของคำ รวมทั้งทำการปรับปรุงอัลกอริทึมที่เป็นการตัดคำแบบเดิมที่เป็นแบบ Recursive มาเป็นอัลกอริทึมที่เป็นแบบ Iterative แทนซึ่งจะสามารถทำการตรวจสอบได้ทุกตำแหน่งของประโยค โดยมีโครงสร้างข้อมูลแบบ Stack ช่วยในการเก็บตำแหน่ง คำในการตัดแบ่งคำ

วัตถุประสงค์ของงานวิจัย

1. สามารถทำการตัดคำที่พบในพจนานุกรมในประโยคได้รวดเร็ว
2. สามารถเพิ่มเติมคำได้รวดเร็ว
3. สามารถค้นหาคำคล้ายได้สะดวก เมื่อพบคำที่ผิด
4. ทำการศึกษาข้อผิดพลาดของ Software Word และนำมาปรับปรุงในงานวิจัยนี้ รวมทั้งเปรียบเทียบความเร็วในการตัดคำด้วย
5. ใช้หน่วยความจำอย่างมีประสิทธิภาพ แต่จะเห็นที่ความเร็วเป็นหลัก

โครงสร้างของวิทยานิพนธ์

โครงสร้างของวิทยานิพนธ์ ประกอบด้วย 5 ส่วน คือ บทที่ 2 กล่าวถึง การค้นหาคำในพจนานุกรมตามโครงสร้างข้อมูลแบบต่างๆ , บทที่ 3 กล่าวถึง แนวทางการออกแบบพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย , บทที่ 4 กล่าวถึง การตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย, บทที่ 5 กล่าวถึง ผลการทดสอบการตัดแบ่งคำภาษาไทย และการเปรียบเทียบกับ Word , บทที่ 6 สรุปผลและข้อเสนอแนะ

บทที่ 2

การค้นหาคำในพจนานุกรมตามโครงสร้างข้อมูลแบบต่างๆ

<u>โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์</u>

โครงสร้างข้อมูลจะหมายถึงการจัดรูปแบบและระเบียนของข้อมูลให้ทราบถึงขอบเขตซึ่งจะนำมาอยู่ในหน่วยความจำ ดังนั้นการออกแบบโครงสร้างข้อมูลที่ดีจะทำให้เกิดการเอื้ออำนวยต่อการประมวลผลของคอมพิวเตอร์ ดังนี้

ก. ทำให้การใช้หน่วยความจำมีประสิทธิภาพ
ข. สามารถนำข้อมูลที่ต้องการใช้มาประมวลผลได้อย่างรวดเร็ว
ค. สามารถทำการค้นหาคำได้รวดเร็ว
ง. สามารถปรับปรุงเพิ่มคำในพจนานุกรมง่ายและพัฒนาได้อย่างมีประสิทธิภาพ

ในโครงสร้างข้อมูลในการค้นหา (Search) ที่สามารถใช้กับพจนานุกรมอิเล็กทรอนิกส์ได้มีหลายแบบ โดยแต่ละแบบจะมีความแตกต่างกันในเรื่องขนาดและความเร็วในการค้นหา ซึ่งในที่นี้จะขอพูดรายละเอียดของแต่ละโครงสร้างข้อมูลอย่างคร่าวๆ จนถึงโครงสร้างข้อมูลที่ได้นำมาใช้ในงานวิจัยที่เลือกใช้อย่างละเอียดโดยงานวิจัยนี้จะเลือกใช้โครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing)

1. โครงสร้างข้อมูลแบบลำดับ (Sequential List) เป็นเทคนิคที่มีการค้นหาคำที่ง่ายที่สุด ซึ่งในการเก็บคำไม่จำเป็นต้องมีการเรียงลำดับคำในพจนานุกรม โดยการอ่านคำจากพจนานุกรมแล้วทำการเปรียบเทียบคำนั้นว่าตรงกับคำที่ต้องการหาหรือไม่ โดยทำการไล่ทีละ 1 คำจนกว่าจะพบคำที่ต้องการหา เวลาที่ใช้ในการค้นหาอย่างช้า (worst case) ของอัลกอริทึมเป็น O (n) [24] อัลกอริทึมของ Sequential Search

```
SU1 : [เริ่มต้น]        i = 0;
SU2 :[จบหรือไม่]        IF  i >= n
                        { break; }          ( ไม่พบ )
SU3 :[เปรียบเทียบ]      IF  KEY = LIST[i]
                        { break; }          ( พบ )
SU4 :[Loop]             i = i + 1;
                        ไป Step SU2;
```

โดยที่ LIST เป็น อะเรย์ที่เก็บคำในพจนานุกรม
n เป็น จำนวนคำทั้งหมดในพจนานุกรม
KEY เป็น คำที่ต้องการค้นหา

2. โครงสร้างข้อมูลแบบเรียงลำดับ (Sorted List) จะเป็นโครงสร้างที่มีการค้นหาเหมือน Sequential Search โดยมีการเรียงลำดับคำในพจนานุกรมมาแล้ว

อัลกอริทึมของ Sorted List

```
SO1  :[เริ่มต้น]        i = 0;
SO2  :[จบหรือไม่]       IF  i >= n
                        { break; }          ( ไม่พบ )
SO3  :[เปรียบเทียบ]     IF  KEY = LIST[i]
                        { break; }          ( พบ )
SO4  :                  IF  KEY < LIST[i]
                        { break; }          ( ไม่พบ )
SO5  :[Loop]            i = i + 1;
                        ไป Step SO2;
```

3. โครงสร้างข้อมูลแบบดรรชนี (Indexed List) เป็นการเพิ่มความเร็วให้กับโครงสร้างข้อมูลแบบลำดับ โดยจะมีการออกแบบเป็น 2 ส่วน ส่วนแรกจะเก็บตัวอักษรแรกของแต่ละคำในพจนานุกรม ซึ่งจะมีตัวชี้ (Index) คำในพจนานุกรม ส่วนที่สองจะเก็บคำทั้งหมดในพจนานุกรม และทำการค้นหาคำที่ต้องการจากการพิจารณาตัวอักษรแรกของคำที่ต้อง

การหาก่อน เพื่อเปรียบเทียบกับส่วนแรกที่เก็บก่อน เมื่อพบจะทราบขอบเขตคำที่ต้องการหาและทำการเปรียบเทียบต่อในส่วนที่สอง ทำให้จะไม่เสียเวลาเหมือนการค้นหาคำแบบ Sequential Search ซึ่งสามารถศึกษาโครงสร้างแบบนี้ ได้จากผลงานวิจัยเรื่อง " การตรวจสอบตัวสะกดด้วยคอมพิวเตอร์ " ของอาจารย์ ยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ [9]

4. โครงสร้างข้อมูลแบบเรียงลำดับที่มีการค้นหาแบบทวิภาค[24] (Binary Search) การค้นหาคำแบบนี้จะเป็นการแก้ข้อเสียของการค้นหาแบบซีเควนเชียล เพราะถ้าหากคำที่ต้องการค้นหาพบอยู่หลังสุดจะต้องเสียเวลาค้นหานาน แต่การค้นหาแบบทวิภาคนี้จะมีการเรียงคำอยู่แล้ว วิธีการค้นหาจะแบ่งคำในพจนานุกรมเป็น 2 ส่วน แล้วนำคำที่พบในตรงกลางมาเปรียบเทียบกับคำที่ต้องการค้น ถ้าคำที่อยู่ตรงกลางในพจนานุกรมมีค่ามากกว่าแสดงว่าคำที่ต้องการจะอยู่ก่อนคำกลางนี้ จากนั้นนำคำทั้งหมดของครึ่งแรกมาแบ่งเป็น 2 ส่วน นำคำตรงกลางการแบ่งมาเปรียบเทียบ ทำเช่นนี้ไปเรื่อยๆจนกว่าจะพบคำที่ต้องการหรือไม่สามารถหาคำนั้นได้ เวลาที่ใช้ในการค้นหาของอัลกอริทึมเป็น O (log n)

อัลกอริทึมของการค้นหาแบบทวิภาค (Binary Search)

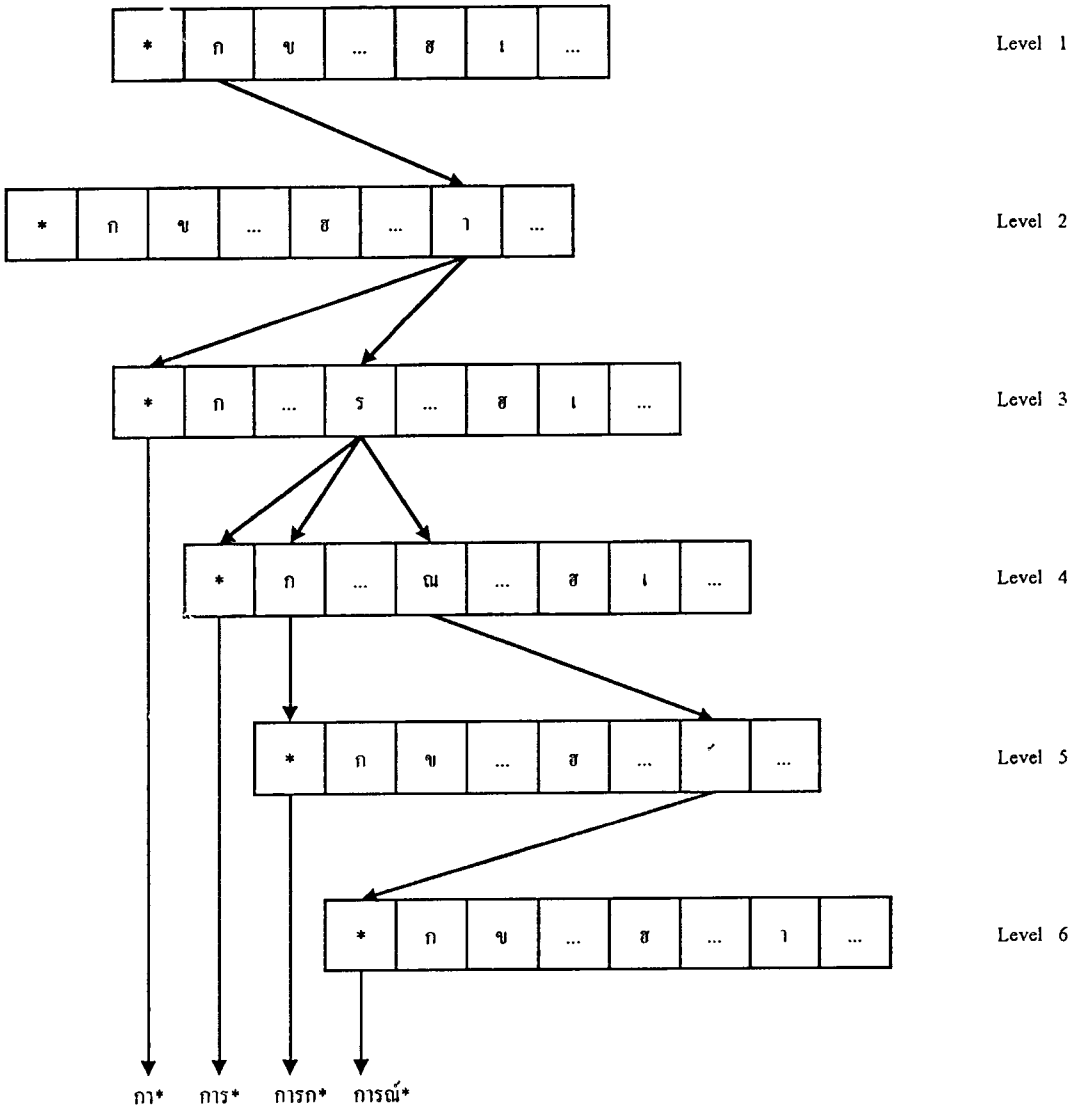
B1	:[เริ่มต้น]	lower = 0; upper = n-1;	
B2	:[จบหรือไม่]	IF upper < lower	
		{ break; }	(ไม่พบ)
B3	:[ค่ากลาง]	i = $\lfloor$(lower + upper) / 2$\rfloor$;	
B4	:[เปรียบเทียบ]	IF KEY = LIST[i]	
		{ break; }	(พบ)
B5	:[เลื่อนค่า]	IF KEY < LIST[i]	
		{ upper = i-1; }	
		ELSE	
		{ lower = i+1; }	
B6	:[Loop]	ไป Step B2	

5. โครงสร้างข้อมูลแบบต้นไม้ (Tree Structure) เป็นโครงสร้างที่มีระดับ (Hierarchical Structure) ซึ่งจะเก็บคำในพจนานุกรมเป็น node โดยมีการเรียงลำดับค่าของคำต่างๆเป็นระดับขั้น (Level) ในการประยุกต์ใช้กับการเก็บคำในพจนานุกรมทั้งหมดจะพบปัญหา

ทำให้ได้มีการออกแบบโครงสร้างข้อมูลแบบทรัย (Trie Structure) ซึ่งเป็นพื้นฐานมาจากโครงสร้างข้อมูลแบบต้นไม้ โดยมีการเก็บคำแต่ละคำเป็น node ของตัวอักษรที่ประกอบกันเป็นคำ ซึ่งจะมีการเชื่อมโยงกันกับตัวอักษรถัดไปของคำ ซึ่งการเก็บอักษรของ node จะมีการเก็บแบบ multiarray ดังภาพที่ 2.1 จะทำให้มีการค้นหาคำที่รวดเร็วกว่าการเก็บโครงสร้างแบบต้นไม้ เวลาที่ใช้ในการค้นหาของอัลกอริทึมเป็น O (log n) [24]

ตัวอย่างการเก็บคำในโครงสร้างข้อมูลแบบทรัย

คำว่า กา, การ, การก, การณ์, ...



ภาพที่ 2.1 แสดงโครงสร้างข้อมูลแบบทรัย (Trie Structure)

โครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing)

เป็นวิธีการค้นหาที่รวดเร็วซึ่งมีเวลาที่ใช้ในการค้นหาของอัลกอริทึมเป็น O (n) [24] โดยวิธีแฮชชิ่งนี้เป็นเทคนิคในการสร้างตารางโดยที่มีค่าคีย์มาช่วยในการค้นหาข้อมูล ซึ่งทำได้ง่ายและรวดเร็ว การแฮชชิ่งจะใช้วิธีการเปลี่ยนค่าของข้อมูลเป็นค่าของแอดเดรสโดยอาศัยฟังก์ชั่น ที่เรียกว่า " แฮชชิ่งฟังก์ชั่น " (Hashing Function หรือ Key - to - Address Transformation) ดังนั้นการกระจายของข้อมูลจะขึ้นกับการเลือกใช้ฟังก์ชั่นที่เหมาะสม แต่ในบางครั้งการแปลงค่าของข้อมูลจะมีค่าแอดเดรสออกมาตรงกัน ซึ่งเรียกว่า " การชนกัน " (Collision) ส่วนประกอบที่สำคัญของโครงสร้างข้อมูลแฮชชิ่ง คือ

1. Hash Table
2. Hash Function

การสร้างฟังก์ชั่นแฮชชิ่ง

1. Modular Division

$$H(x) = x \bmod M$$

โดยที่ x เป็นค่าของข้อมูล

M เป็นขนาดของ Table (prime number)

2. Polynomial Addressing

ถ้ามีค่าของข้อมูลเป็นตัวอักษรในคำที่มีความยาว L ของคำ วิธีการนี้จะสามารถทำงานได้อย่างมีประสิทธิภาพ โดยมีการคำนวณตาม ASCII ของตัวอักษร

ซึ่งจะเขียนเป็นสูตรทั่วไปได้ ดังนี้

$$H(x) = base^{(L-1)}*x_1 + base^{(L-2)}*x_2 + ... + x_L$$

L - ความยาวของคำ

base - เป็นค่าตัวอักษรทั้งหมดในตาราง ASCII

เช่น การ

H(x) = base^2*'ก' + base^1*'า' + 'ร'

โดยจากสูตรทั่วไปของ Polynomial นี้สามารถเขียนเป็นโครงสร้างได้ดังนี้

$h = x_1;$

For (I=2; I <= L; I++)

{ h_i = (h * base + x_i) mod Size ; }

size - ขนาดของ Table Hashing (Prime Number)

3. Multiplication

$H(x) = floor(Size * (k * x - floor(k * x)))$

โดยที่ k เป็นค่า $0 < k < 1$

Size เป็นขนาด Table

<u>การแก้ปัญหาคอลลิชัน</u> [3] (Collision)

จากการใช้แฮชชิ่งฟังก์ชั่น จะพบการชนกันของข้อมูลอยู่บ้าง ทำให้เกิดปัญหาในการค้นหาข้อมูล ซึ่งจะมีการแก้ไขดังนี้

1. Open - Addressing
2. Chaining

<u>วิธีการ Open - Addressing</u> (ภาพที่ 2.2)

1. Linear Probing เป็นวิธีที่ง่ายที่สุด คือ ถ้าพบว่ามีการชนกันของข้อมูลเกิดขึ้นให้บวกค่าคงที่เข้ากับ H(x)

$H(x) = H(x) + C$

อัลกอริทึมของ Linear Probing มีดังนี้

HI1	: [ค่าเริ่มต้น]	loc = H(KEY);
HI2	: [การตรวจสอบ]	IF (HLIST [loc] == EMPTY)
		THEN HLIST[loc] = KEY ; (เก็บค่า)
		ไปที่ HI1;
HI3	: [การเลื่อน]	loc = loc + 1;
		IF (loc == M)
		THEN loc = 0;
HI4	: [Loop]	ไปยังขั้น HI2 ;

โดยที่

loc	เป็นค่าตำแหน่งการ Hash
H(KEY)	Hashing Function
KEY	ข้อมูล
HLIST	อะเรย์ในการเก็บค่า
M	ขนาด Table

2. Random Probing

$$H_{i+1}(x) = (H_i(x) + C) \bmod M$$

อัลกอริทึมของ Random Probing มีดังนี้

1. คำนวณค่าเริ่มต้น
2. ถ้าค่าเริ่มต้นที่คำนวณได้เป็นค่าแอดเดรสที่ยังว่างให้ใส่ค่าลงไปในแอดเดรสแล้วออกจากโพรซีเจอร์นี้
3. ทำซ้ำถึงข้อ 5 จนกว่าจะพบแอดเดรสว่าง
4. คำนวณหาแอดเดรสที่ว่างจากสูตร Random Probing
5. ถ้าทุกแอดเดรสถูกตรวจสอบแล้วว่าไม่ว่างให้เขียน Overflow ออกจากโพรซีเจอร์
6. ใส่ค่าในแอดเดรสนี้ แล้วออกจากโพรซีเจอร์

3. Double Hashing เป็นการใช้ฟังก์ชั่น 2 ฟังก์ชั่น โดยมีฟังก์ชั่นหนึ่งเป็นฟังก์ชั่นอิสระ และอีกฟังก์ชั่นจะเป็นการเลือกจากค่า Random Probing

ตัวอย่างเช่น สมมติว่ามีฟังก์ชั่น H_1 ที่ $H_1(x_1) = H_1(x_2) = i$ เมื่อ $x_1 \neq x_2$ และถ้ามีฟังก์ชั่น G_1 ที่ $G_1(x_1) \neq G_1(x_2)$ เมื่อ $x_1 \neq x_2$ เราจะใช้ $G_1(x_1)$ หรือ $G_1(x_2)$ เป็นค่าพารามิเตอร์ c (ค่าคงที่) เมื่อนำไปใช้กับวิธี random probing จะได้ค่าที่แตกต่างกันเมื่อ H_1 และ G_1 ต่างก็เป็นอิสระกัน

ถ้าให้ $H_1(x) = x$ MOD m และ $G_1(x) = (x$ MOD $(m-2)) + 1$ เมื่อ x เป็นค่าคีย์ และ m เป็นขนาดตาราง ถ้า $x = 75$ และ $m = 11$ จะได้ $H_1(75) = 9$ และ $G_1(75) = 4$

เมื่อนำมาประยุกต์ใช้กับ random probing โดยใช้สูตร

$$H_{i+1}(x) = (H_i(x) + c) \text{ MOD } 11$$

เมื่อ $i = 1$ จะได้ $H_1(75) = 9$ และ $c = G_1(75) = 4$

เมื่อค่า $i = 1, 2, 3, \ldots$ จะได้ผลลัพธ์ของ $H_i(75)$ ดังนี้

9, 2, 6, 10, 3, 7, 0, 4, 8, 1, 5

และสมมติว่ามีคีย์ถัดไปเป็น 42

$H_1(42) = 9$ (ค่าเดียวกับ $H_1(75)$ แต่ $G_1(42) = 7 = c$)

เมื่อใช้วิธี random probing

$$H_{i+1}(x) = (H_i(x) + c) \text{ MOD } 11$$

แทนค่าเริ่มต้น เมื่อ $i = 1$ จะได้ $H_1 = 9$ และ $c = 7$

เมื่อค่า $i = 1, 2, 3, \ldots$ จะได้ผลลัพธ์ของ $H_i(42)$ ดังนี้

2, 5, 1, 8, 4, 0, 7, 3, 10, 6, 2

จะเห็นว่าถ้าเกิดการชนกัน $H_1(75) = H_1(42) = 9$ ถ้าใช้วิธีนี้จะได้ค่าที่แตกต่างกัน

<u>วิธีการ Chaining</u> (ภาพที่ 2.3)

เป็นวิธีที่นิยมกันมาก โดยมีลิงค์ลิสต์มาช่วย ซึ่งจะแบ่งการ Chaining เป็น 2 ส่วน คือ ส่วนแรกจะมีขนาดคงที่ โดยจะเป็นลักษณะของอะเรย์ ส่วนที่สองจะเป็นพอย์นเตอร์ในการชี้ค่าที่มีแอดเดรสเดียวกันกับค่าของแอดเดรสข้อมูลแรก ซึ่งในงานวิจัยนี้ใช้วิธีนี้ในการแก้ไขปัญหาการชนกันของข้อมูล

ตัวอย่างการแก้ปัญหาการชนกันของคำ

ธาตุ, นักร, ปัชชุน, ยอก, เลียบเคียง, ทระนง, สบู่, นิบาต

การคำนวณแอดเดรสผ่านฟังก์ชั่นแฮชชิ่ง

$$H_{i+1} = ((H_i * 67) + X_{i+1}) \bmod 32761$$

H - ค่า Hashing

X - Ascii ของตัวอักษรถัดไปในคำ

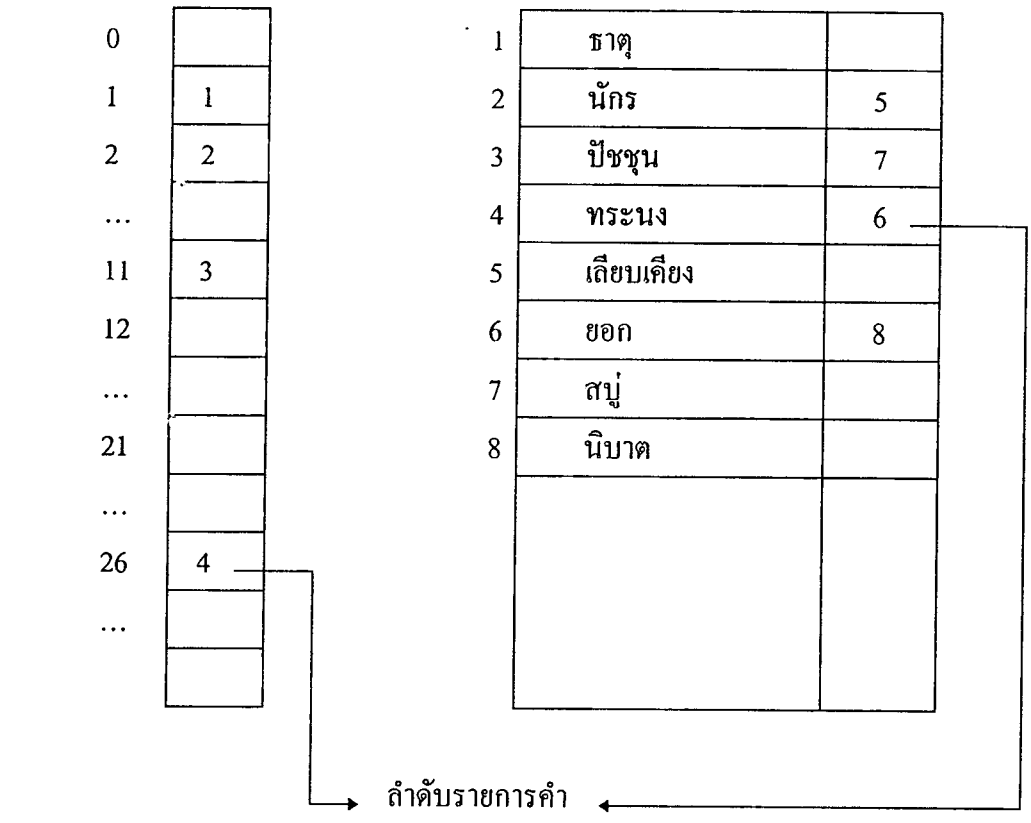
คำนวณรอบตามตัวอักษรที่ประกอบเป็นคำจะพบว่า

คำว่า	ธาตุ	มีแอดเดรสเป็น	1
คำว่า	นักร, เลียบเคียง	มีแอดเดรสเป็น	2
คำว่า	ปัชชุน, สบู่	มีแอดเดรสเป็น	11
คำว่า	ทระนง, นิบาต, ยอก	มีแอดเดรสเป็น	26

ค่าแอดเดรส			จำนวน Prob
1	ธาตุ		1
2	นักร		1
3	เลียบเคียง		2
...			
11	ปัชชุน		1
12	สบู่		2
...			
26	ทระนง		1
27	นิบาต		2
28	ยอก		3
...			

ภาพที่ 2.2 การแก้ปัญหาการชนกันโดยวิธี Open Addressing

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง



ภาพที่ 2.3 แสดงการแก้ปัญหาการชนกันโดยวิธี Chaining

31664

บทที่ 3

แนวทางการออกแบบพจนานุกรมอิเล็กทรอนิกส์

พจนานุกรมอิเล็กทรอนิกส์ เป็นแหล่งข้อมูลในด้านภาษาศาสตร์ ที่จำเป็นสำหรับการประมวลผลภาษาธรรมชาติ [7] พจนานุกรมที่จัดเก็บลงในคอมพิวเตอร์ จึงมีบทบาทที่สำคัญเพิ่มมากขึ้น ในงานประมวลผลดังนั้นการเก็บพจนานุกรมอิเล็กทรอนิกส์ จึงมีบทบาทเป็นอย่างมากในการค้นหาคำ ส่วนรายละเอียดของพจนานุกรมก็จะแตกต่างกันออกไปตามแต่การใช้งานต่าง ๆ

<u>ประเภทของพจนานุกรมอิเล็กทรอนิกส์</u> อาจจะแบ่งเป็น

1. ประเภทผู้ใช้งาน
   1.1. พจนานุกรมอเนกประสงค์ (General Dictionary) เช่น พจนานุกรมฉบับราชบัณฑิตยสถาน ปทานุกรมนักเรียน
   1.2. พจนานุกรมศัพท์เทคนิค (Technical Dictionary) เก็บข้อมูลคำศัพท์เฉพาะสาขา เช่น ทางการแพทย์, วิศวกรรม, คอมพิวเตอร์
   1.3. พจนานุกรมเล็กซิคอน (User Specific Lexicon) เก็บข้อมูลศัพท์เฉพาะงานใดงานหนึ่งโดยเฉพาะ เช่น พจนานุกรมสำหรับงานแปลภาษา

2. ตามคู่ภาษา (Classify by Language Pairs)
   2.1. พจนานุกรมที่มีคู่ภาษาเพียงภาษาเดียว (Mono - Lingual Dictionary) เช่น พจนานุกรมฉบับราชบัณฑิตยสถาน พจนานุกรมของลองแมน
   2.2 พจนานุกรมที่มีคู่ภาษา 2 ภาษา (Bilingual Dictionary) เช่น พจนานุกรมไทย - อังกฤษ

2.3. พจนานุกรมที่มีคู่ภาษามากกว่า 2 ภาษา (Multilingual Dictionary)

3. เนื้อหาของข้อมูลที่บันทึก

3.1. พจนานุกรม (Dictionary)

3.2. เล็กซิคอน (Lexicon)

3.3. พจนานุกรมคำพ้อง (Thesaurus)

3.4. สารานุกรม (Encyclopedia)

4. ตามกลุ่มผู้ใช้

4.1. พจนานุกรมสำหรับมนุษย์ (Dictionary for Human)

4.2. พจนานุกรมสำหรับเครื่องใช้ (Dictionary for Computer)

<u>รายละเอียดข้อมูลในพจนานุกรมอิเล็กทรอนิกส์</u>

พจนานุกรมอิเล็กทรอนิกส์มีมากมายหลายประเภทขึ้นอยู่กับวัตถุประสงค์ของผู้ใช้งานเพราะฉะนั้นเนื้อหาในการเก็บรายละเอียดก็จะแตกต่างกันไป ดังนั้นหากว่ามีพจนานุกรมอิเล็กทรอนิกส์ที่สามารถเก็บรายละเอียดข้อมูลทุกอย่างที่สามารถนำไปใช้งานหลายๆด้านก็จะทำให้พจนานุกรมนั้นมีประสิทธิภาพในการใช้งานสูง สำหรับรายละเอียดที่สามารถเก็บในพจนานุกรมอิเล็กทรอนิกส์ได้ เราอาจจะแบ่งเป็นการเก็บของ

ก. คำหรือหน่วยคำ

ข. ข้อมูลระดับคำ และการแบ่งแยกคำและพยางค์

ค. ข้อมูลเกี่ยวกับการผันของคำ

ง. รายละเอียดเกี่ยวกับการผันของคำ

จ. การออกเสียง

ฉ. ชนิดของคำ

ช. ความถี่ที่ใช้และความสำคัญของคำ

ซ. ข้อมูลเกี่ยวกับลักษณะของคำ

ฌ. โครงสร้างของไวยกรณ์

ญ. คำพ้องความหมาย

ฎ. การเชื่อมโยงในคำ และ ไวยกรณ์

ฯลฯ

รายละเอียดของพจนานุกรมอิเล็กทรอนิกส์ที่กล่าวมาในตอนต้นนี้ เราอาจจะมีการแบ่งรายละเอียดในการเก็บได้อีกมาก ขึ้นอยู่กับการใช้งานที่จำเป็นในทางการประมวลผลข้อมูลด้านภาษาศาสตร์

<u>ลักษณะของคำไทย</u>[14]

คำในภาษาไทยนั้นมีการแบ่งเป็น 2 ประเภท คือ

ก. คำโดด หมายถึง คำเดียวที่มีความหมายในตัวเอง เช่น ยืน เดิน วิ่ง

ข. คำประสม หมายถึง คำที่เกิดจากการนำคำโดดมารวมกันแล้วเกิดคำใหม่ที่มีความหมาย เช่น เรือใบ น้ำตก

นอกจากนี้การแบ่งคำในภาษาไทยยังแตกต่างจากการแบ่งคำในภาษาอังกฤษ เนื่องจากวิธีการเขียนประโยคภาษาไทยนั้นเขียนติดต่อโดยไม่มีช่องว่างดังการเขียนประโยคในภาษาอังกฤษ ซึ่งมีการเขียนที่มีช่องว่างระหว่างคำ ทำให้ไม่สามารถทราบถึงตำแหน่งสิ้นสุดคำในประโยคภาษาไทยที่แน่นอน เพราะฉะนั้นการเลือกโครงสร้างข้อมูลสำหรับการเก็บพจนานุกรมและวิธีปฏิบัติการจะต้องนำลักษณะคำไทยและการแบ่งคำมาเป็นหลักการพิจารณาด้วย

<u>โครงสร้างของพยางค์ในภาษาไทย</u>

คำไทยส่วนใหญ่จะเป็นคำพยางค์เดียว (Monosyllable) แต่จะมีบางคำที่มีหลายพยางค์ เราสามารถแบ่งโครงสร้างพยางค์ของภาษาไทยได้ 3 กลุ่ม คือ

1. โครงสร้างของคำพยางค์เดียว จะแบ่งเป็น 2 ลักษณะ

ว

1.1) พ (พ) ส พ เช่น จน , เป็น , กรุง ฯลฯ

ว

1.2) พ (พ) ส : (พ) เช่น อย่า , ปลา , กลัว

โดยที่ พ ------ พยัญชนะ

(พ) ------ พยัญชนะควบกล้ำ อาจจะมีหรือไม่ก็ได้

ส ------ สระเสียงสั้น

ส : ------ สระเสียงยาว

ว ------- วรรณยุกต์

2. โครงสร้างของคำสองพยางค์ จะเป็นการประสมของคำพยางค์เดียว จะมีรูปแบบ 2 ลักษณะ คือ

2.1. พวกที่มีพยางค์แรกเป็นสระเสียงสั้น ได้แก่ สนั่น, สะกิด, กะพริบ ฯลฯ

2.2. พวกที่มีพยางค์แรกเป็นสระเสียงยาวหรือมีตัวสะกด ได้แก่ กากี, มานพ, กันดาร, วันที่ ฯลฯ

3. โครงสร้างของคำสามพยางค์ขึ้นไป มักจะเป็นคำประสมหรือคำที่ยืมมาจากภาษาอื่น ได้แก่ ไวโอลิน, วิทยาลัย

<u>ชนิดของหน่วยคำในภาษาไทย</u> แบ่งเป็น 2 ชนิด

1. หน่วยคำอิสระ (Free morpheme) คือ หน่วยคำที่ปรากฏได้ตามลำพังในประโยค หรือปรากฏร่วมกับหน่วยคำอื่นบางหน่วยที่ไม่อาจปรากฏได้ตามลำพัง

ตัวอย่าง เช่น ตาย

1. แมวตายแล้ว

2. เราไม่กลัวแม้แต่ความตาย

จะเห็นว่าคำว่า " ตาย " ในประโยคที่ 1 จะสามารถอยู่ได้ตามลำพัง ส่วนประโยคที่ 2 จะต้องมีการประสมคำระหว่าง " ความ " กับ " ตาย "

2. หน่วยคำไม่อิสระ (Bound morpheme) คือ หน่วยคำที่ปรากฏตามลำพังไม่ได้ในประโยค จะแบ่งเป็น 5 ประเภท

2.1 หน่วยหน้าศัพท์ (prefix) คือ หน่วยคำไม่อิสระที่เกิดหน้าหน่วยคำอื่นเสมอ

เช่น

/ การ - / , / ความ - / , / นัก - / , / ชาว - / , / อธิ - / , / อภิ - / , / อนุ - /

2.2 หน่วยหน้าศัพท์คู่ (double prefix) คือ หน่วยคำไม่อิสระที่มีรูปและเสียงเหมือนกัน 1 คู่ เช่น

/ กระ - กระ - /	จุ๋มจิ๋ม	กระจุ๋มกระจิ๋ม
/ ขะ - ขะ - /	มอมแมม	ขะมอมขะแมม
/ อี - อี - /	เก้กัง	อีเก้อีกัง

2.3 หน่วยกลางศัพท์ (infix) คือ หน่วยคำไม่อิสระที่ปรากฏกลางหน่วยคำอื่น เช่น

/ - ำ - /	เสร็จ	สำเร็จ
/ - ำ น - /	เกิด	กำเนิด

2.4 หน่วยท้ายศัพท์ (suffix) คือ หน่วยคำไม่อิสระที่เกิดท้ายหน่วยคำอื่น เช่น

/ - การ /	กรรมการ	กิจการ
/ - ภาพ /	อิสรภาพ	มรณภาพ
/ - ศาสตร์ /	วิทยาศาสตร์	ภูมิศาสตร์

2.5 หน่วยเสริมสร้อย (chining reduplication) คือ หน่วยคำไม่อิสระที่จะปรากฏ หน้าหรือ หลังหน่วยคำอื่น เช่น

/ - หนังหา /	หนังสือหนังหา
	ฯลฯ

<u>การวิเคราะห์ตัวอักษรในภาษาไทย</u>

จากการวิเคราะห์คำไทยในพจนานุกรมฉบับราชบัณฑิตยสถาน จะสามารถแบ่งตัวอักษรไทยได้ ดังนี้ (รหัส ASCII ของตัวอักษรไทยในภาคผนวก ก)

1. กลุ่มพยัญชนะ 44 ตัว คือ ก ข ฃ ----- ฮ
2. กลุ่มสระ 17 ตัว คือ ะ ั า ำ ิ ี ึ ื ุ ู เ แ โ ใ ไ ฤ ฦ

3. กลุ่มวรรณยุกต์ 4 ตัว คือ ◌่ ◌้ ◌๊ ◌๋

4. กลุ่มตัวเลข 10 ตัว คือ 0 1 2 3 4 5 6 7 8 9

5. กลุ่มอักษรอื่น ๆ คือ ◌็ ◌์ ฯ ๆ ๅ ( ) " ! ? : ; , .

การวิเคราะห์ของตัวอักษรพยัญชนะในภาษาไทย แบ่งเป็น 5 กลุ่ม คือ

1. ฉ ผ ฝ ฮ จะเป็นตัวอักษรแรกของพยางค์
2. ห ภ ฆ ฟ ซ ฑ ส่วนมากจะเป็นตัวอักษรแรกของพยางค์
3. อ ว ร จะสามารถเป็นทั้งตัวพยัญชนะและตัวสระในคำ ซึ่ง ถ้าตัว ร จะเป็นรูป รร จะถือว่าเป็นสระของคำ
4. ศ ณ ญ ษ ฐ ฎ ฏ ฒ ฬ ฌ ส่วนใหญ่จะเป็นตัวสะกด
5. ก ข ค ง จ ช ค ภ ถ ท น ย ป พ ม ย ล ส จะสามารถเป็นได้ทั้งตัวสะกดและเป็นตัวอักษรนำ

การวิเคราะห์ของตัวอักษรสระในภาษาไทย แบ่งเป็น

1. เ แ ไ โ ใ จะเป็นตัวนำของคำ
2. ะ ำ จะเป็นตัวสุดท้ายของคำและไม่ต้องการตัวสะกด
3. ◌ั ◌ี ◌ึ ◌ื จะเป็นสระที่ต้องมีตัวสะกด
4. า ◌ิ ◌ี ◌ุ ◌ู จะเป็นสระที่มีตัวสะกดหรือไม่มีตัวสะกดก็ได้
5. ฤ จะเป็นสระของพยางค์บางคำ

การวิเคราะห์ของตัวอักษรวรรณยุกต์ในภาษาไทย แบ่งเป็น

1. เ แ โ ไ ใ จะอยู่ข้างหน้าคำโดยมีพยัญชนะตาม และมีวรรณยุกต์ด้วยก็ได้
2. สระ ฤ ◌็ และ ◌์ จะไม่มีวรรณยุกต์ตาม
3. ◌ั ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู จะมีวรรณยุกต์ตามหลัง
4. ะ า ำ อ และ ว จะมีวรรณยุกต์อยู่หน้าของกลุ่มคำนี้

<u>การออกแบบพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย</u>

แนวทางการออกแบบโครงสร้างข้อมูลสำหรับการจัดเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยในวิทยานิพนธ์นี้ ได้พิจารณาจากวิธีการค้นหาคำศัพท์จากพจนานุกรมลักษณะคำไทย และโครงสร้างของภาษาไทย โดยโครงสร้างข้อมูลนี้จะเป็นโครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing) โดยมีการค้นหาคำศัพท์ เป็นแบบแฮชชิ่งด้วย ซึ่งโครงสร้างข้อมูลแฮชชิ่ง เป็นโครงสร้างที่รวดเร็วต่อการสืบค้นหาคำ และสะดวกต่อการเพิ่มเติมคำศัพท์ใหม่ ๆ ในพจนานุกรมอิเล็กทรอนิกส์ ซึ่งโครงสร้างข้อมูลแบบแฮชชิ่ง มักจะพบการชนกัน (Collision) ของข้อมูล ถ้ามีฟังก์ชั่นแฮชชิ่งที่ไม่ดีในงานวิจัยได้นำฟังก์ชั่นแฮชชิ่ง แบบ Polynomial Addressing มาทำการคำนวณโดยใช้ฟังก์ชั่นของ Horner ช่วยในการคำนวณ โดยงานวิจัยนี้จะทำการแก้ไขปัญหาการชนกันของข้อมูลด้วยวิธีการของ Chaining

<u>วิธีการปฏิบัติการกับข้อมูล</u> แบ่งเป็น 2 ประเภท

1. การปฏิบัติการแบบพลวัต (Dynamic Method) เป็นวิธีการเก็บข้อมูลในรูปตารางที่มีการเปลี่ยนแปลงได้ เช่น ต้นไม้แบบทวิภาค (Binary Tree) ต้นไม้แบบบีพลัส (B^+ - Tree) โครงสร้างข้อมูลแบบทรัย (Trie)

2. การปฏิบัติการแบบสถิต (Static Method) เป็นวิธีการจัดเก็บข้อมูลที่ไม่มีการเปลี่ยนแปลง หรือ ตารางตายตัว เช่น การหาแฮชชิ่งแบบสมบูรณ์ (Perfect Hashing) ทรัยแบบลดขนาด (Compressed Trie)

วิธีการปฏิบัติการกับโครงสร้างข้อมูลทั้ง 2 แบบ มีข้อดีและข้อเสียที่แตกต่างกันโดยข้อดีแบบพลวัต คือ โครงสร้างข้อมูลมีการเปลี่ยนแปลงในขณะที่มีการปรับเปลี่ยนของข้อมูล ส่วนข้อดีของแบบสถิต คือ โครงสร้างมีขนาดเล็กและสามารถค้นหาได้รวดเร็ว สำหรับวิทยานิพนธ์นี้ได้นำข้อดีทั้ง 2 มาผสมปรับปรุงเพื่อจะเก็บคำในพจนานุกรมต่อไป

<u>โครงสร้างในการเก็บคำในพจนานุกรมอิเล็กทรอนิกส์</u>

การเก็บโครงสร้างข้อมูลแบบแฮชชิ่งนี้ จะแบ่งเป็น 2 ส่วน คือ

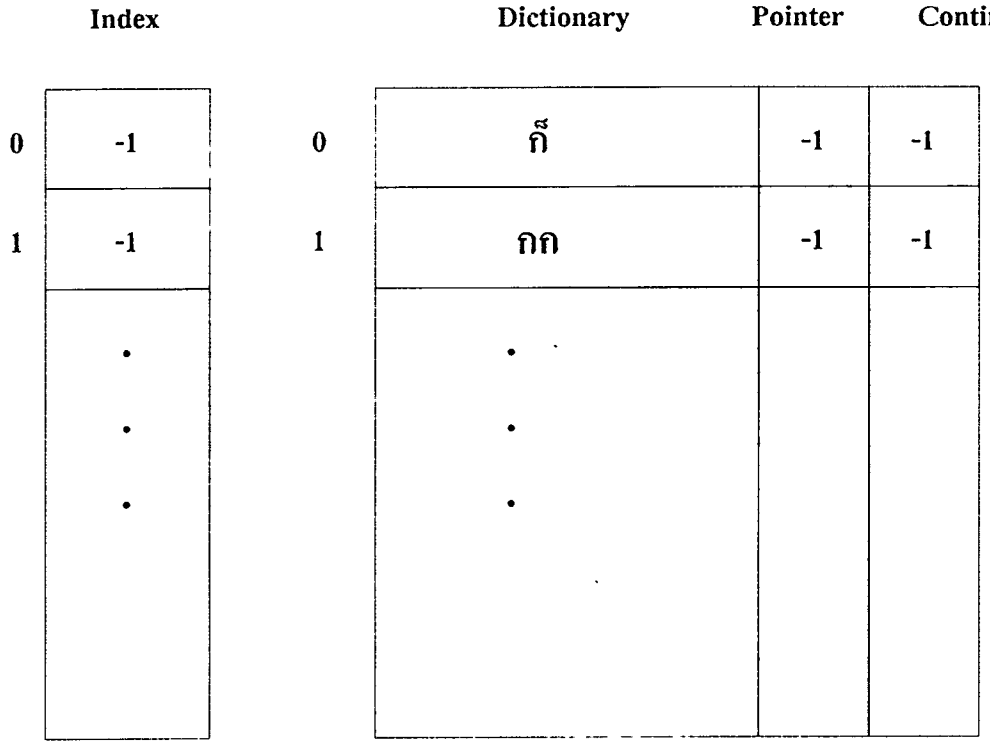
1. ส่วนเก็บ Index จะเก็บค่าที่คำนวณแอดเดรส
2. ส่วนเก็บคำในพจนานุกรม โดยส่วนนี้จะมีโครงสร้างดังภาพที่ 3.1 จะแบ่งเป็น
   2.1 คำศัพท์
   2.2 ตัวชี้ Address ในการ Link เมื่อพบว่ามีคำแอดเดรสในการ hash เดียวกัน
   2.3 ตัวบ่งบอกการมีคำต่อ (Continue Search)

คำ	Pointer	Continue
การ	-1	-1

ภาพที่ 3.1 โครงสร้างการเก็บคำของพจนานุกรม

<u>ขั้นตอนการออกแบบและการสร้างพจนานุกรมอิเล็กทรอนิกส์</u>

ขั้นที่ 1 จะนำคำทั้งหมดในพจนานุกรม มาเก็บไว้ในโครงสร้างภาพที่ 3.1 จะได้โครงสร้างก่อนที่จะนำมาคำนวณผ่าน Hashing Function ดังภาพที่ 3.2 โดยที่ค่าเริ่มต้นของการเก็บโครงสร้างจะมีค่าเท่ากับ -1



ภาพที่ 3.2 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing Structure)

ขั้นที่ 2 นำคำในพจนานุกรมมาทำการคำนวณผ่าน hashing function ทีละคำใน พจนานุกรม โดยวิธีการ Polynomial Addressing ซึ่งใช้หลักการของ Horner เป็นตัวช่วยในการคำนวณ ซึ่งจะขอยกตัวอย่างการคำนวณ ดังนี้ โดยมีฟังก์ชั่นที่ใช้

$$h_{i+1} = [(h_i * base) + X_{i+1}] \% size$$

ซึ่งจะคำนวณตามจำนวนตัวอักษรที่ประกอบในคำนั้น

โดยที่ h แทน ค่าแอดเดรสที่คำนวณได้ ค่าเริ่มต้นคำเท่ากับ 0

base แทน จำนวนทั้งหมดของรหัส ASCII ซึ่งตัวอักษรทั้งหมดมี 256 ตัว

X แทน ค่ารหัส ASCII (ภาคผนวก ก) ของตัวอักษรในคำ

size แทน ขนาดของตาราง Hashing ซึ่งจะต้องเป็น prime number ในวิทยานิพนธ์ นี้จะให้ค่าเท่ากับ 32761

ตัวอย่าง คำว่า " ก็ "

X_0 = " ก " = 161

X_1 = " ็ " = 231

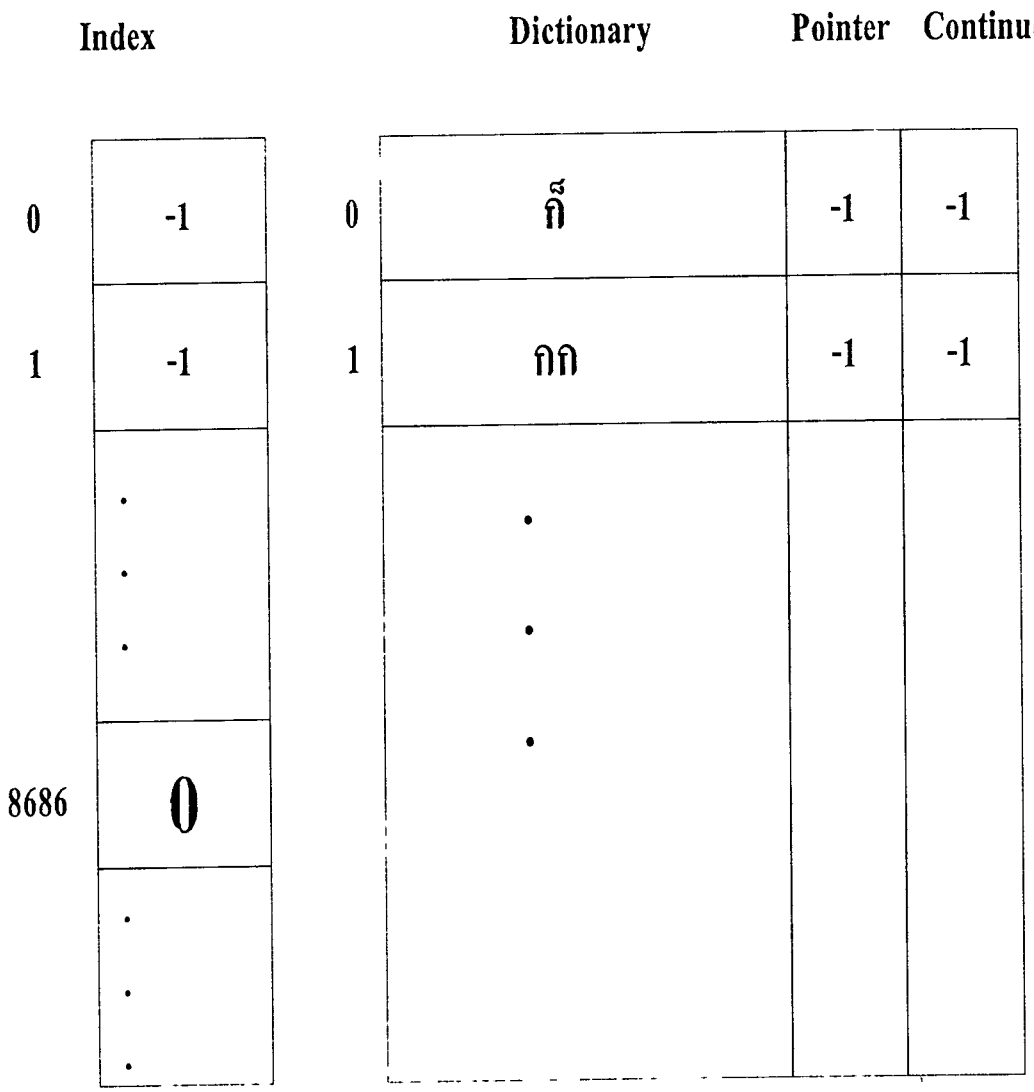
รอบที่ 1 h = ((0 * 256) + 161) % 32761 = 161

รอบที่ 2 h = ((161 * 256) + 231) % 32761 = 8686

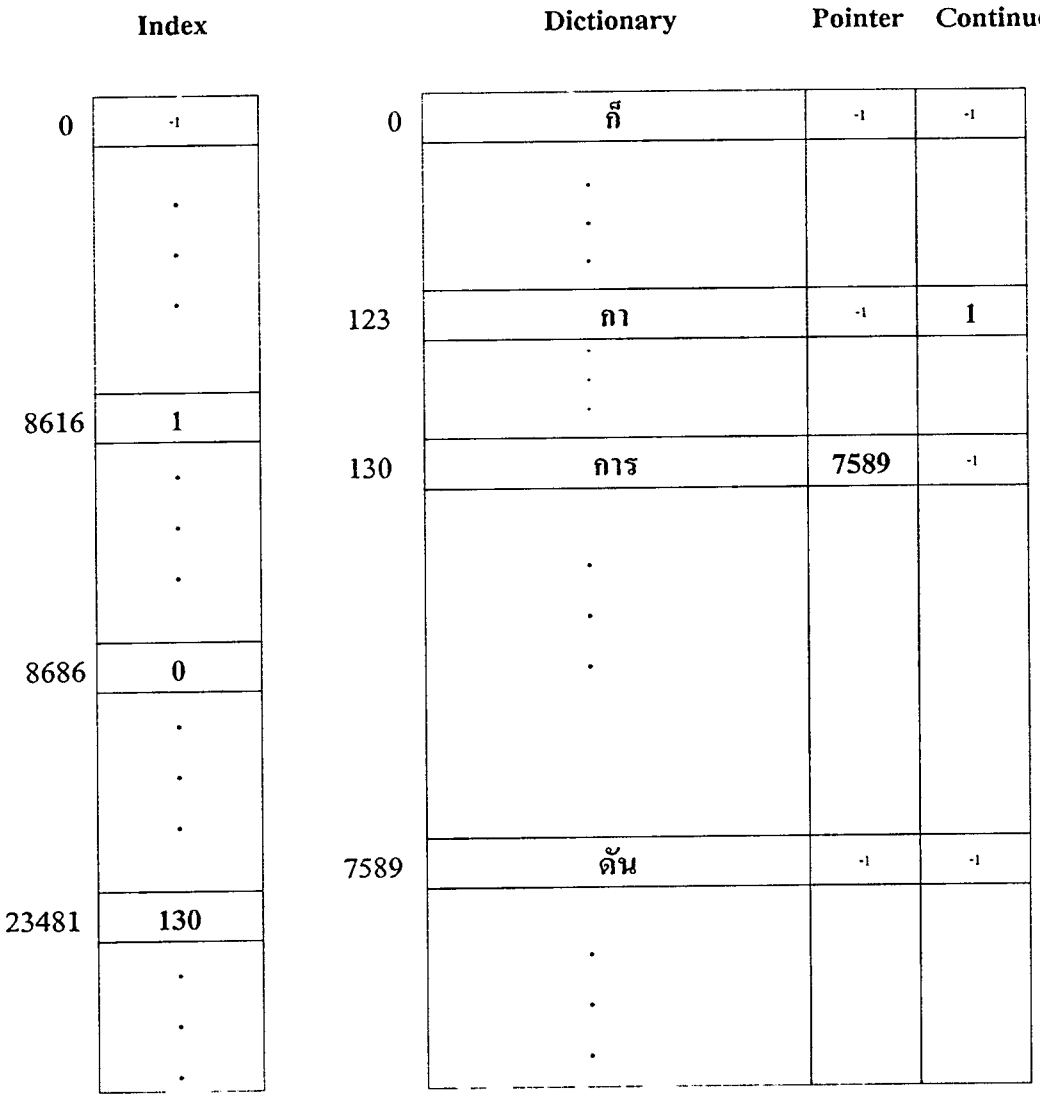
ดังนั้น คำว่า " ก็ " จะมีค่าแอดเดรส เท่ากับ 8686

ขั้นที่ 3 เมื่อได้ค่าแอดเดรสของคำแล้ว จะทำการเปลี่ยนค่าในตาราง Index ดังภาพที่ 3.3 โดยการตรวจสอบก่อนว่าค่าในตาราง Index เป็นค่าที่เท่ากับ -1 หรือไม่ ถ้าเท่ากับ -1 ก็ให้ทำการเปลี่ยนค่าได้ แต่ถ้าไม่เท่ากับ -1 ก็ให้นำการค้นหาไปยังส่วนที่เป็นแอดเดรสของคำในตารางดรรชนี (Index) เพื่อให้หาคำที่เป็นคำในพจนานุกรมที่มีค่าแอดเดรสของคำสุดท้ายที่มีค่าเป็น -1 ของโครงสร้างดังภาพที่ 3.1 ซึ่งค่านั้นจะต้องเท่า -1 ถ้าไม่เท่ากับ -1 ให้ทำการค้นหาต่อไปเรื่อยๆ จนกว่าค่าจะเท่ากับ -1 จึงเปลี่ยนค่าในฟิลด์นั้น ให้เป็นค่าแอดเดรสของคำใหม่

ขั้นที่ 4 จะทำการคำนวณในขั้นตอนที่ 2-3 จนกว่าคำในพจนานุกรมจะหมด โดยค่าฟิลด์ที่ 3 จากภาพที่ 3.1 จะเปลี่ยนค่าเป็น 1 เมื่อพบว่ามีคำที่ทำการคำนวณนั้นมีตัวอักษรต่อจากคำๆ นั้น ซึ่งจะแสดงให้เห็นดังภาพที่ 3.4 โดยคำว่า " กา " จะมีค่า " การ " ซึ่งเป็นตัวที่มีอักษรต่อจากคำว่า " กา " โดยจะเปลี่ยนค่าฟิลด์ของ Continue เป็นค่าเท่ากับ 1 (มีตัวต่อ)



ภาพที่ 3.3 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing Structure)



ภาพที่ 3.4 แสดงการเก็บคำของโครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing Structure)

<u>ผลของการเก็บโครงสร้างพจนานุกรมแบบ Hashing</u>

จากการเก็บโครงสร้างพจนานุกรม Hashing นี้ จะพบว่ามีการชนกัน (Collision) ของคำในการคำนวณ คำจากคำทั้งหมดในพจนานุกรม 32,381 คำ พบว่าค่าแอดเดรสที่ชนกันมากที่สุดมีจำนวน 6 Collision จำนวน 3 คำ

โดยจะมีค่าเฉลี่ยในการค้นหาคำในพจนานุกรมเท่ากับ 0.50 คำต่อครั้ง ดังตารางที่ 3.1 โดยให้วิทยานิพนธ์ นี้จะทำการหาผลที่เกี่ยวกับการใช้พจนานุกรม โดยแบ่งเป็นขนาดของคำในพจนานุกรม สถิติเกี่ยวกับการใช้ของตัวอักษรที่ประกอบขึ้นเป็นคำ และพร้อมทั้งทำการเขียนโปรแกรมเก็บค่าสถิติอื่น ๆ ที่เกี่ยวข้องกับพจนานุกรมด้วย

ตารางที่ 3.1 จำนวนการเกิด Collision

จำนวน Collision	จำนวนคำ	เปอร์เซ็นต์ในพจนานุกรม
ไม่มีCollision	20540	63.43
1 Collision	8484	26.20
2 Collision	2566	7.92
3 Collision	635	1.96
4 Collision	134	0.41
5 Collision	19	0.07
6 Collision	3	0.01
รวม	32381	
ค่าเฉลี่ย	0.50	

ตารางที่ 3.2 ตารางแสดงขนาดตัวอักษรที่ประกอบเป็นคำ

ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์ %
1	1	0.003
2	471	1.455
3	2522	7.789
4	4031	12.449
5	4890	15.101
6	5597	17.285
7	4660	14.391
8	3408	10.525
9	2177	6.723
10	1489	4.598
11	934	2.884
12	695	2.146
13	426	1.316
14	336	1.038
15	215	0.664
16	150	0.463
17	90	0.278
18	77	0.238
19	49	0.151

ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์ %
20	38	0.117
21	41	0.127
22	18	0.056
23	10	0.031
24	10	0.031
25	9	0.028
26	8	0.025
27	11	0.033
28	3	0.009
29	2	0.006
30	1	0.003
31	2	0.006
32	3	0.009
33	2	0.006
34	2	0.006
35	1	0.003
36	1	0.003
37	0	0
38	1	0.003

ตารางที่ 3.3 แสดงการแจกแจงการใช้ตัวอักษรเป็นตัวแรก

ก	= 2735	ป	= 1691	◌็	= 0
ข	= 902	ผ	= 432	◌่	= 0
ฃ	= 0	ฝ	= 134	◌้	= 0
ค	= 964	พ	= 1139	◌๊	= 0
ฅ	= 0	ฟ	= 139	◌๋	= 0
ฆ	= 19	ภ	= 334	เ	= 3706
ง	= 122	ม	= 1123	แ	= 1043
จ	= 820	ย	= 494	โ	= 1023
ฉ	= 180	ร	= 944	ใ	= 178
ช	= 681	ฤ	= 46	ไ	= 606
ซ	= 199	ล	= 910	ๅ	= 0
ฌ	= 4	ฦ	= 6	ๆ	= 0
ญ	= 13	ว	= 904	◌์	= 0
ฎ	= 1	ศ	= 318	◌	= 0
ฏ	= 0	ษ	= 10	◌	= 0
ฐ	= 20	ส	= 2080	◌	= 0
ฑ	= 4	ห	= 1329	◌	= 0
ฒ	= 0	ฬ	= 0	◌	= 0
ณ	= 2	อ	= 1951		
ด	= 455	ฮ	= 61		
ต	= 1307	ฯ	= 1		
ถ	= 318	ะ	= 0		
ท	= 986	◌ั	= 0		
ธ	= 239	า	= 0		
น	= 954	◌ำ	= 0		
บ	= 854	◌ิ	= 0		

ตารางที่ 3.4 แสดงการแจกแจงการใช้ตัวอักษรเป็นตัวสุดท้าย

ก	= 2022	น	= 3535	◌ะ	= 0
ข	= 69	บ	= 897	า	= 2581
ฃ	= 0	ป	= 159	◌ำ	= 430
ค	= 198	ผ	= 7	◌ิ	= 516
ฅ	= 0	ฝ	= 1	◌ี	= 1055
ฆ	= 33	พ	= 249	◌ึ	= 7
ง	= 3527	ฟ	= 75	◌ื	= 0
จ	= 326	ภ	= 40	◌ุ	= 273
ฉ	= 7	ม	= 1819	◌ู	= 293
ช	= 180	ย	= 1881	เ	= 0
ซ	= 20	ร	= 1422	แ	= 0
ฌ	= 2	ฤ	= 3	โ	= 0
ญ	= 166	ล	= 703	ใ	= 0
ฎ	= 10	ฦ	= 0	ไ	= 0
ฏ	= 39	ว	= 1058	ๅ	= 4
ฐ	= 52	ศ	= 238	ๆ	= 116
ฑ	= 17	ษ	= 140	◌็	= 2
ฒ	= 14	ส	= 302	◌่	= 546
ณ	= 225	ห	= 58	◌้	= 468
ด	= 1411	ฬ	= 16	◌๊	= 13
ต	= 560	อ	= 863	◌๋	= 40
ถ	= 83	ฮ	= 4	◌์	= 1958
ท	= 265	ฯ	= 1		
ธ	= 84	ะ	= 1298		

ตารางที่ 3.5 การแจกแจงความถี่ของตัวอักษร

ก	= 11052	น	= 11369	◌ั	= 7950
ข	= 2557	บ	= 4237	า	= 14858
ฃ	= 0	ป	= 3944	◌ำ	= 1679
ค	= 3430	ผ	= 944	◌ิ	= 6481
ฅ	= 0	ฝ	= 297	◌ี	= 4562
ฆ	= 177	พ	= 3464	◌ึ	= 623
ง	= 7600	ฟ	= 466	◌ื	= 1837
จ	= 2879	ภ	= 1082	◌ุ	= 4075
ฉ	= 545	ม	= 7946	◌ู	= 2032
ช	= 2446	ย	= 7098	เ	= 8006
ซ	= 618	ร	= 12825	แ	= 2484
ฌ	= 43	ฤ	= 314	โ	= 2122
ญ	= 831	ล	= 6517	ใ	= 652
ฎ	= 133	ฦ	= 6	ไ	= 1468
ฏ	= 216	ว	= 6172	ๅ	= 22
ฐ	= 298	ศ	= 1330	ๆ	= 130
ฑ	= 229	ษ	= 928	◌็	= 822
ฒ	= 59	ส	= 5470	◌่	= 5431
ณ	= 1226	ห	= 4880	◌้	= 6430
ด	= 4934	ฬ	= 104	◌๊	= 228
ต	= 5626	อ	= 8791	◌๋	= 204
ถ	= 1007	ฮ	= 176	◌์	= 2206
ท	= 3930	ฯ	= 2		
ธ	= 1202	ะ	= 5445		

ตารางที่ 3.6 การแจกแจงความถี่ของตัวอักษรที่ใช้มากตามลำดับ

พยัญชนะ

ที่	ตัวอักษร	จำนวน	ที่	ตัวอักษร	จำนวน
1	ร	12825	23	ณ	1226
2	น	11369	24	ธ	1202
3	ก	11052	25	ภ	1082
4	อ	8791	26	ถ	1007
5	ม	7946	27	ผ	944
6	ง	7600	28	ษ	928
7	ย	7098	29	ญ	831
8	ล	6517	30	ซ	618
9	ว	6172	31	ฉ	545
10	ต	5626	32	ฟ	466
11	ส	5470	33	ฐ	298
12	ด	4934	34	ฝ	297
13	ห	4880	35	ฑ	229
14	บ	4237	36	ฎ	216
15	ป	3944	37	ฆ	177
16	ท	3930	38	ฮ	176
17	พ	3464	39	ฏ	133
18	ค	3430	40	ฬ	104
19	จ	2879	41	ฒ	59
20	ข	2557	42	ฌ	43
21	ช	2446	43	ฃ	0
22	ศ	1330	44	ฅ	0

สระ

ที่	ตัวอักษร	จำนวน	ที่	ตัวอักษร	จำนวน
1	า	14858	10	◌ู	2032
2	เ	8006	11	◌ี	1837
3	◌ั	7950	12	◌ำ	1679
4	◌ิ	6481	13	ไ	1468
5	ะ	5445	14	◌็	623
6	◌่	4562	15	ใ	652
7	◌ุ	4075	16	ฤ	314
8	แ	2484	17	ฦ	6
9	โ	2122			

วรรณยุกต์และตัวอักษรอื่นๆ

ที่	ตัวอักษร	จำนวน	ที่	ตัวอักษร	จำนวน
1	◌้	6430	5	◌๊	228
2	◌	5431	6	◌๋	204
3	◌์	2206	7	ๆ	130
4	◌๋	822	8	ๅ	22

ตารางที่ 3.7 ตารางแสดงจำนวนคำที่ตัวอักษรประกอบเป็นคำมากที่สุด 10 อันดับ

ลำดับ	ขนาดตัวอักษร		ลำดับ	ขนาดตัวอักษร	
1	6	ตัวอักษร	6	3	ตัวอักษร
2	5	ตัวอักษร	7	9	ตัวอักษร
3	7	ตัวอักษร	8	10	ตัวอักษร
4	4	ตัวอักษร	9	11	ตัวอักษร
5	8	ตัวอักษร	10	12	ตัวอักษร

อัลกอริทึมในการค้นหาคำ

การค้นหาคำจากโครงสร้างข้อมูลแบบ Hashing นั้น จะเป็นการตรวจสอบว่าคำที่ต้องการค้นหามีในโครงสร้างพจนานุกรมแฮชชิ่งหรือไม่มี ซึ่งมีขั้นตอนการค้นหา ดังนี้

1. นำคำที่ต้องการค้นหามาทำการคำนวณผ่าน Hashing Function

2. เมื่อได้ค่าแอดเดรส ให้นำค่าแอดเดรสนั้นมาพิจารณาในส่วนของ Index Table ก่อน เมื่อได้ค่าตำแหน่งของคำจาก Index Table ให้มาพิจารณาที่ Dictionary Table โดยดูจากค่าตำแหน่งใน Index Table

3. เมื่อพิจารณาคำที่ได้จากตำแหน่งของ Dictionary Table นำคำมาทำการเปรียบเทียบกับคำที่ต้องการค้นหา เมื่อพบคำนั้นก็จะหยุดค้นหา และเมื่อไม่ใช่คำที่ต้องการค้นหาก็จะทำการพิจารณาในฟิลด์ที่ 2 ของโครงสร้างพจนานุกรมในภาพที่ 3.1 ว่ามี Link (พิจารณาจากค่า -1) ต่อหรือไม่ ถ้าไม่มี Link ต่อก็ทำการหยุดค้นหา เพราะคำที่ต้องการไม่พบในพจนานุกรม แต่ถ้ามี Link (ไม่เท่ากับ -1) ต่อก็ให้ทำการเปรียบเทียบคำต่อไปกับคำที่ต้องการค้นหา จนกว่าจะพบหรือสิ้นสุด Link (เป็น -1 หรือไม่)ที่เชื่อมต่อ

ตัวอย่างแสดงขั้นตอนในการค้นหา

สมมติว่าต้องการทราบว่า คำว่า " การ " เก็บไว้ในพจนานุกรมอิเล็กทรอนิกส์จัดเก็บไว้ในโครงสร้างข้อมูลแบบ Hashing หรือไม่ รายละเอียดมีดังนี้

1. นับจำนวนตัวอักษรของคำที่จะค้นหา

" การ " = 3

2. ทำการคำนวณค่า Address โดยทำการคำนวณตามรอบที่ประกอบเป็นคำ

รอบที่ 1 : h = ((0 * 256) + 161 (" ก ")) % 32761 = 161

รอบที่ 2 : h = ((161 * 256) + 210 (" า ")) % 32761 = 8665

รอบที่ 3 : h = ((8665 * 256) + 195 (" ร ")) % 32761 = 23448

3. นำค่าแอดเดรสที่คำนวณได้ มาพิจารณาใน Index Table เพื่อดูตำแหน่งของคำใน พจนานุกรมในภาพที่ 3.4 ซึ่งจะได้ตำแหน่งคำที่ 130 ในพจนานุกรม นำคำนั้นมาทำการเปรียบเทียบทีละตัวอักษร จะพบว่าคำว่า " การ " พบในพจนานุกรม

<u>อัลกอริทึม ในการค้นหา</u>

SH 1 : [นับความยาวของคำ] len = strlen (KEY) ;

SH 2 : [คำนวณค่า Hash] For (i = 0 , i <= len , i ++)

{ h (KEY) = [(h * 256) + X_{i+1}] % 32761 ; }

SH 3 : [check if found ใน Index]

If (INDEX [h] == -1)

{ break ; } (ไม่พบ)

If (INDEX [h] != -1)

{ loc = INDEX [h] ; } (พบว่ามีคำแต่อาจจะใช่หรือไม่)

SH 4 : [check if found]

If (DICTIONARY [loc] == KEY)

{ break ; } (พบคำที่ต้องการหา)

If (DICTIONARY [loc] != KEY)

{ loc = DICTIONARYPOINTER [loc] ; }

SH 5 : [Loop] Go to SH 4

โดยที่ len = เป็นจำนวนตัวอักษรที่ประกอบกันเป็นคำ

strlen = เป็นฟังก์ชั่นนับความยาวของคำ ในภาษา C

h (Key) = ค่าแอดเดรสที่คำนวณของคำ

INDEX = ตารางของคำ Index ในการชี้ไปยังตำแหน่งใน DICTIONARY

DICTIONARY = ตารางในการเก็บคำในพจนานุกรม แบ่งเป็น 3 ส่วน

1. ส่วนคำ (DICTIONARYDATA)
2. ส่วนตัวชี้ (DICTIONARYPOINTER)
3. ส่วนตัวต่อจากคำ (DICTIONARYCONTINUE)

loc = ตำแหน่งของคำ

KEY = คำที่ต้องการหา

X = ค่าของตัวอักษรในตาราง ASCII

<u>การออกแบบโครงสร้างพจนานุกรมในการค้นหาคำคล้าย</u>

ในการออกแบบการเก็บคำในพจนานุกรมนั้นมีความสำคัญ แต่ในการเก็บโครงสร้างของพจนานุกรมในการหาคำคล้ายก็มีคำสำคัญมากเหมือนกัน เมื่อในการค้นหาไม่สามารถค้นหาคำที่ต้องการได้ ก็ควรจะมีคำที่คล้ายกับคำที่ต้องการหาขึ้นมาให้ผู้ใช้ได้ดู เพราะคำที่ต้องการค้นหานั้นอาจจะมีการพิมพ์ที่ผิดพลาดเกิดขึ้น โดยโครงสร้างการเก็บจะแตกต่างกับโครงสร้างพจนานุกรมแบบ Hashing เดิมเล็กน้อย ซึ่งโครงสร้างใหม่นี้จะเก็บตาราง Index ของแอดเดรส เป็น 2 ส่วน คือ ส่วนทางขวาและส่วนทางซ้าย โดยมีขั้นตอนการสร้างดังนี้

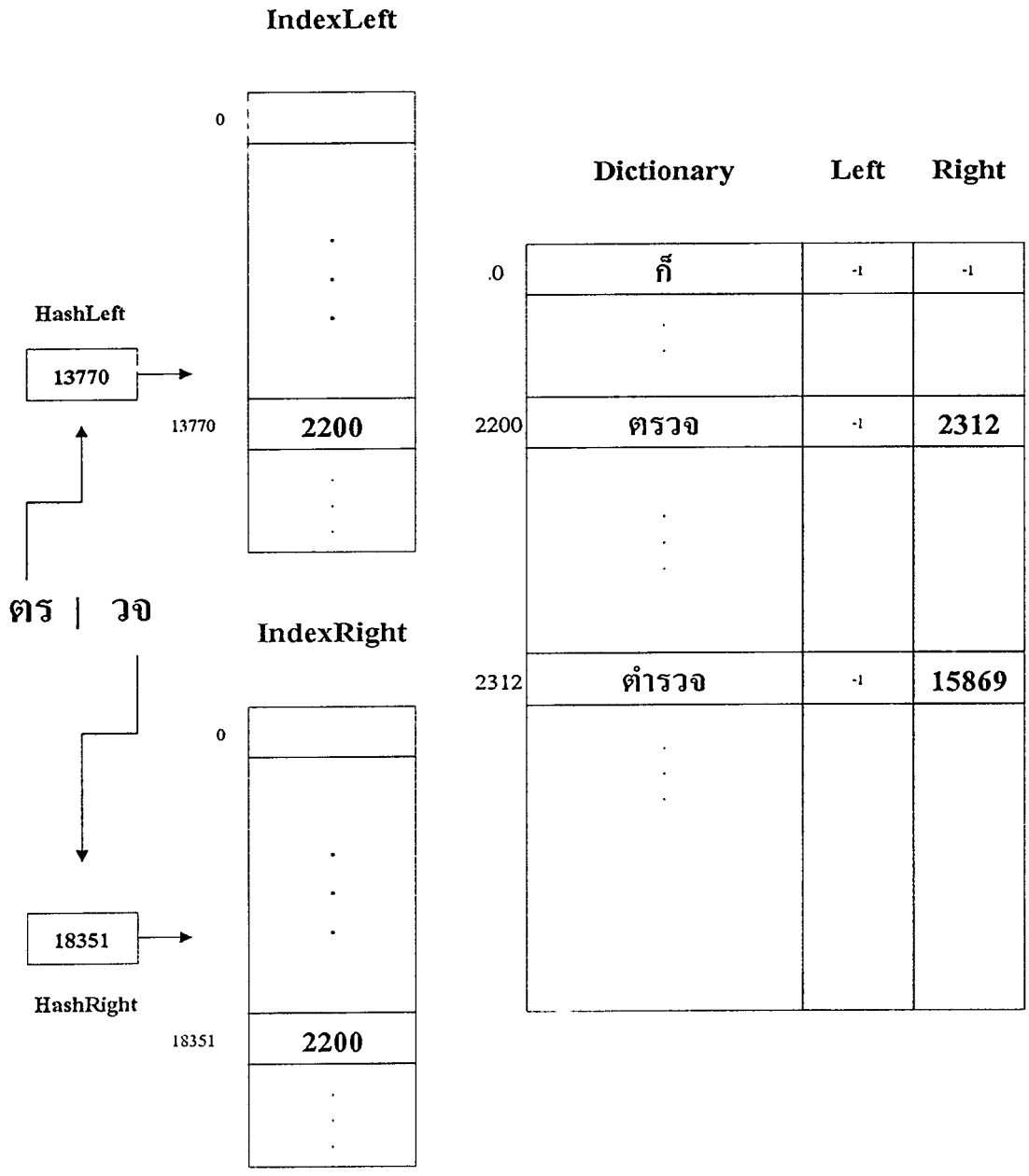
1. นำคำในพจนานุกรมมาทำการคำนวณผ่าน Hash Function ทีละคำ โดยที่นับความยาวของตัวอักษรในคำ แล้วทำการแบ่งครึ่งคำนั้นมาคำนวณทีละครั้ง (ถ้ามีจำนวนตัวเป็นคี่จะแบ่งเป็นทางซ้ายมากกว่าทางขวา) มาคำนวณ

2. เมื่อได้ค่าแอดเดรสของครึ่งคำทางซ้ายจะนำมาใส่ในตาราง Index Left ซึ่งเป็น Index ทางซ้าย และค่าแอดเดรสทางขวาของคำ จะไปใส่ค่าในตาราง Index Right ซึ่งเป็น Index ทางขวา

3. เมื่อพบค่าแอดเดรสที่เหมือนกันจะทำการสร้าง Link เชื่อมโยงกับที่ส่วนทางท้ายของโครงสร้างคำที่มีแอดเดรสเดียวกัน

<u>ตัวอย่างการสร้างพจนานุกรมคำคล้าย</u>

สมมติมี คำว่า " ตรวจ "

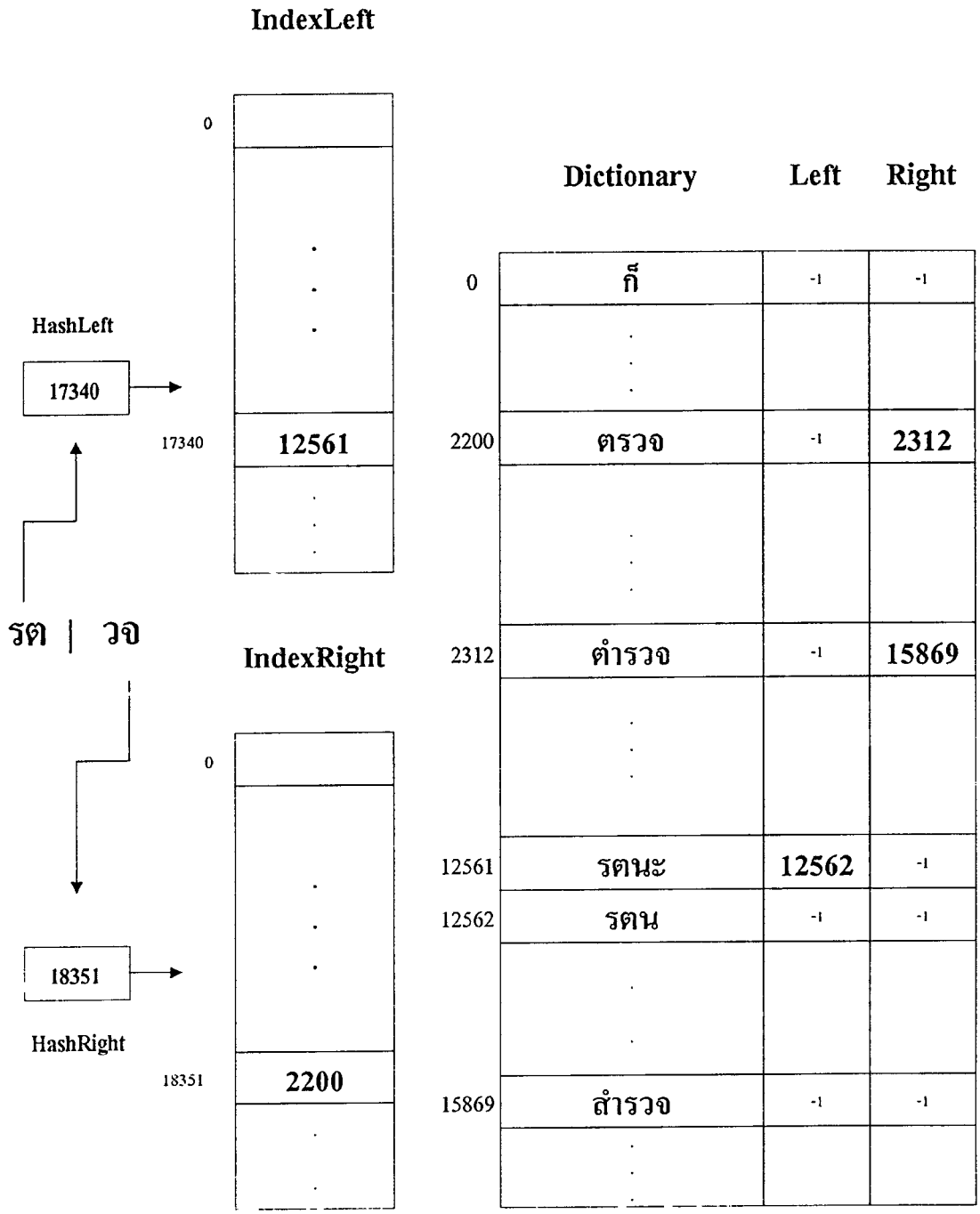


ภาพที่ 3.5 รูปแสดงวิธีการสร้างพจนานุกรมคำคล้าย

จากภาพที่ 3.5 การสร้างพจนานุกรมแฮชชิ่งหาคำคล้ายทำโดยการนำคำมาแบ่งครึ่งสร้างดรรชนีคำทางซ้ายและดรรชนีคำทางขวา ซึ่งจากตัวอย่างจะนำคำว่า ตรวจ มาแบ่งครึ่งสร้างดรรชนีคำว่า ตร เป็นดรรชนีทางซ้าย และสร้างคำว่า วจ เป็นดรรชนีคำทางขวาเพื่อใช้ค้นหาคำต่อไป

อัลกอริทึมในการค้นหาคำคล้ายในพจนานุกรม

วิธีการหาคำคล้ายจะใช้ก็ต่อเมื่อพบว่ามีคำที่ไม่พบในพจนานุกรมเกิดขึ้น โดยเราจะแบ่งคำที่ต้องการหาเป็น 2 ส่วน เพื่อหาคำคล้าย ทั้งจากคำที่คล้ายกับในส่วนซ้าย และคำที่คล้ายกันในส่วนขวาของคำที่ต้องการหา ดังตัวอย่างดังภาพที่ 3.6 เช่น คำว่า " รตวจ "



ภาพที่ 3.6 แสดงวิธีการค้นหาคำคล้ายในพจนานุกรม

จากภาพที่ 3.6 จะได้คำที่คล้ายกับคำว่า " รตวจ " จะมีคำว่า รตนะ , รตน , ตรวจ , ตำรวจ, สำรวจ เพื่อให้สามารถทำการแก้ไขต่อไป

บทที่ 4

การตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

ในบทนี้จะกล่าวถึงการตัดแบ่งคำของภาษาไทย โดยมีวิธีการตัดแบ่งคำได้ 2 แบบ โดยแต่ละวิธีจะมีข้อดีและข้อเสียที่ต่างกัน โดยจะกล่าวถึงวิธีการตัดคำทั้ง 2 แบบ พร้อมทั้งนำพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ได้ทำการออกแบบได้มาช่วยในการตัดแบ่งด้วย ซึ่งการตัดแบ่งคำภาษาไทยจะมีวิธีดังนี้

1. การตรวจสอบทางอักษรวิธี จะทำการตรวจสอบการสะกดคำโดยการตรวจสอบพยัญชนะ, สระ, วรรณยุกต์, การันต์ ว่าอยู่ในตำแหน่งที่ควรจะเป็นหรือไม่ในลำดับของคำและกฎต่าง ๆ เช่น

พยัญชนะ
- ฃ ฅ ไม่ใช้แล้วจึงไม่ควรปรากฏในคำ
- ฌ ญ ฏ ฒ ณ ฬ ไม่ปรากฏเป็นอักขระต้นของคำนอกจากมีพวกสระนำมาก่อน ยกเว้น ณ เพียงคำเดียว

ฯลฯ

สระ
- สระหน้าจะต้องตามด้วยพยัญชนะเสมอ
- สระหลังจะต้องมีพยัญชนะหรือวรรณยุกต์นำหน้าเท่านั้น

ฯลฯ

วรรณยุกต์
- ต้องตามด้วยตัวสะกดหรือสระหลังเท่านั้น
- เป็นอักขระแรกของคำไม่ได้

ฯลฯ

การันต์
- ต้องอยู่ท้ายสุดของคำเท่านั้น ยกเว้นคำที่มาจากภาษาอังกฤษ
- ต้องตามหลังพยัญชนะบางตัวหรือสระอิ เท่านั้น

ฯลฯ

ข้อดีของวิธีการตรวจสอบทางอักษรวิธี จะเป็นไปอย่างรวดเร็ว เนื่องจากเป็นการตรวจสอบกฎที่ตั้งเอาไว้แล้ว ทำการเปรียบเทียบมีจำนวนน้อยครั้งกว่าการตรวจสอบคำในพจนานุกรมเนื่องจากฐานข้อมูลที่จะต้องใช้ในวิธีการนี้ คือ กฎต่าง ๆ ในการตรวจสอบ ย่อมจะมีน้อยกว่าจำนวนคำทั้งหมดในฐานข้อมูลพจนานุกรม แต่มีข้อเสียคือ ความแม่นยำในการตรวจสอบจะน้อย เนื่องจากจะพบคำที่ผิดแต่เป็นไปตามกฎ เช่น งาง วอล เป็นต้น นอกจากนี้ในการปฏิบัติจะทำได้ยาก

2. การตรวจสอบกับพจนานุกรม คือ นำคำไปเปรียบเทียบกับคำในพจนานุกรมว่าพบหรือไม่ ข้อดี จะมีความแม่นยำสูง จึงแน่ใจว่าคำที่ผ่านการตรวจสอบทุกคำมีความหมายแน่นอน นอกจากนี้การปฏิบัติยังทำได้ง่าย แต่ความเร็วในการตรวจสอบจะช้า เพราะต้องเปรียบเทียบกับคำในพจนานุกรม นอกจากนี้ยังไม่สามารถตรวจสอบคำที่ไม่มีในพจนานุกรมได้

<u>การวิเคราะห์โครงสร้างพยางค์ของคำภาษาไทย</u> [4],[14]

การวิเคราะห์คำภาษาไทยโดยทั่วไป จะพบว่าจะมีอยู่ 7 รูปแบบ คือ

1) C(C) $\overset{(T)}{}$ V(S(S))

2) C(C) $\overset{V}{\overset{(T)}{}}$ (S(S))

3) C(C) $\overset{(T)}{\underset{V}{}}$ (S(S))

4) VC(C) $\overset{(T)}{}$ (S(S))

5) VC(C) $\overset{(T)}{}$ V(S(S))

6) VC(C) $\overset{V}{\overset{(T)}{}}$ (S(S))

7) VC(C) $\overset{(T)}{\overset{V}{}}$ V(S(S))

โดยที่

C	-	พยัญชนะ
V	-	สระ
T	-	วรรณยุกต์
S	-	ตัวสะกด
( )	-	อาจจะมีหรือไม่มี

<u>การวิเคราะห์การแบ่งพยางค์ไทยแบบอักษรวิธี</u>

1) การจัดแบ่งกลุ่มของพยางค์ไทย

1.1) กลุ่มตายตัว เช่น สระ ◌ึ จะไม่มีรูปแบบ C◌ึ และ C◌ึSS

1.2) กลุ่มที่คลุมเครือ เช่น สระ ◌ิ จะพบรูปแบบ C◌ิ, C◌ิS และ C◌ิSS ซึ่งจะไม่สามารถทำการแบ่งพยางค์ได้ทันที

2) การกำหนดขอบเขตของคำ มีหลักเกณฑ์ ดังนี้

2.1) พิจารณาขอบเขตคำ ที่เป็นตามหลักอักษรวิธี ยกเว้นคำที่คลุมเครือ

2.2) พิจารณาโดยใช้ขอบเขตของการเป็นตัวหน้าและตัวสุดท้ายของคำ

2.3) ทำการ mark คำที่ไม่อยู่ในหลักเกณฑ์ เพื่อพิจารณาที่หลัง

<u>กฎการตัดคำโดยพิจารณาจากอักษรวิธี</u> [22]

1. กฎของการตัดแบ่งขอบเขตพยางค์ตัวหน้า

1.1) สระ า ◌ิ ◌ี ◌ุ ◌ู ◌ึ ◌ื ◌็ ะ ◌ำ ◌ั และวรรณยุกต์ ◌่ ◌้ ◌๊ ◌๋ จะต้องมีพยัญชนะนำอย่างน้อย 1 ตัว

1.2) สระ เ แ ไ โ จะต้องมีพยัญชนะนำตามหลัง ยกเว้นคำบางคำ เช่น ขโมย

1.3) สระ ใ จะมีพยัญชนะตามเสมอ

1.4) ฉ ผ ฝ ฮ เป็นตัวนำหน้าเสมอ ยกเว้นมีสระ เ แ โ ไ ใ มาก่อน

1.5) ห ส่วนมากจะเป็นตัวนำของคำ ยกเว้น สห มหา คหบดี มหกรรม

1.6) ข้อ 1.1 อาจจะมีพยัญชนะมากกว่า 1 ตัว

2. กฎการตัดแบ่งพยางค์ตัวหลัง

2.1) ศ ณ ญ ษ ฐ ฎ ฏ ฒ ฬ ฌ ส่วนมากเป็นตัวสะกด ยกเว้น ศก ศร ศพ ยก ญวน

2.2) ◌ึ จะพบตัวสะกดเสมอ ยกเว้น รึ, หึ, ฮึ

2.3) ไม้หันอากาศ จะมีตัวสะกด 1 ตัว

2.4) ◌ำ และ ะ จะเป็นตัวสุดท้ายของพยางค์

2.5) ◌์ มักพบเป็นตัวสุดท้ายของคำ ยกเว้น คำจากภาษาอังกฤษ

2.6) วรรณยุกต์ที่อยู่ระหว่าง 2 พยัญชนะ ไม่มีสระจะต้องการสะกดแน่นอน

2.7) วรรณยุกต์ที่ไม่มีสระในพยางค์ และตามด้วย อ จะต้องการ 1 ตัวสะกด

2.8) ◌ิ ◌ี ◌ึ ◌ั จะมีวรรณยุกต์ จะต้องการ 1 ตัวสะกด

2.9) ◌ึ ปรากฏ กับวรรณยุกต์อื่นนอกจาก ◌็ ส่วนมากไม่ต้องการตัวสะกด

2.10) เ◌็◌ แ◌็◌ เ◌ิ◌ จะต้องมี 1 ตัวสะกด

2.11) ◌ัว ◌ัย ◌ัฯ ◌ือ เ◌า เ◌อ เ◌าะ เ◌อะ

เ◌ะ แ◌ะ โ◌ะ เ◌ียะ เ◌ือะ จะไม่มีตัวสะกด ยกเว้น

เ◌า และ เ◌อ

2.12) สระ ไ จะมีในคำภาษาไทย 20 คำ

ผู้ใหญ่หาผ้าใหม่ ให้สะใภ้ใช้คล้องคอ ใฝ่ใจเอาใส่ห่อ มิหลงใหล ใครขอดู
จะใคร่ลงเรือใบดูน้ำใส และ ปลาปู สิ่งใดอยู่ในตู้ มิใช่อยู่ใต้ตั่งเตียง
บ้าใบ้ ถือใยบัว หูตามัวมาใกล้เคียง เล่าท่องอย่าละเลี่ยง 20 ม้วน จำจงดี

3) กฎการตัดแบ่งคำอื่น ๆ

3.1) ในภาษาไทย . จะใช้เป็นคำย่อ สามารถเป็นขอบเขตของคำได้

3.2) ฯ จะพบในคำ ฯลฯ และ ฯ พณ ฯ

3.3) ตัวอักษรย่อเป็นตัวเลขและภาษาต่างประเทศ เป็นการบอกขอบเขตของคำภาษาไทย

<u>การวิเคราะห์การตัดแบ่งคำภาษาไทยโดยการตรวจสอบจากพจนานุกรม</u>

จะมีหลักการ 2 ข้อ ดังนี้

1. การตรวจสอบการสะกดคำทำโดยการเปรียบเทียบกับคำที่ได้ในพจนานุกรม
2. การแยกคำจะแยกคำตามคำยาวที่สุดในพจนานุกรม

ลักษณะที่พบในการตรวจสอบการสะกดคำและการแยกคำมี 5 ลักษณะ ดังนี้

1. คำที่ถูกต้อง และ mark จุดแยกคำได้ เช่น

แปลง่าย เมื่อตรวจสอบถึงคำว่า แปล จะเห็นว่าพบในพจนานุกรม และยังมีคำอื่นอีกที่เป็นไปได้ เช่น แปลง แปร แปล ดังนั้นจึงต้องนำการ mark ไว้ก่อนที่จุดแยกคำ ทีหลัง " ป " และทำการเปรียบเทียบต่อ จะพบว่า แปล ก็เป็นจุดที่สามารถแยกคำได้ ก็ทำการ mark ที่หลัง " ล " ต่อ

2. คำที่ถูกต้อง และแยกคำได้ เช่น

แปลง่าย เมื่อตรวจสอบถึง แปลง จะเห็นว่าพบคำในพจนานุกรม และไม่มีคำอื่นที่เป็นไปได้ ดังนั้น จะทำการแยกคำทีหลัง " ง " ทันที แต่อาจจะผิด เมื่อเปรียบเทียบคำต่อไป

3. คำที่ไม่ถูกต้อง แต่สามารถย้อนกลับได้ เช่น

แปลง่าย เมื่อตัดคำทีหลัง " ง " ไปแล้ว ก็จะเริ่มตรวจสอบคำต่อไป โดยเริ่มจาก แต่พบว่าไม่มีคำที่ขึ้นต้นด้วย =' ในพจนานุกรม ก็ตรวจสอบว่ามีการ mark ไว้ก่อนหน้าหรือไม่ ถ้ามีการ mark จะย้อนกลับไปที่ mark ก่อนหน้า แล้วตรวจสอบใหม่ โดยเริ่มจากตัวอักษร " ง " เปรียบเทียบกับคำในพจนานุกรมต่อ

4. คำที่ถูกต้อง แต่ไม่สามารถ mark หรือแยกคำได้ เช่น

ตลาด เมื่อตรวจสอบถึงตัว ตล จะพบว่า เป็นส่วนของคำที่มีในพจนานุกรมเท่านั้น จึงไม่สามารถทำการ mark หรือ แยกคำได้ ต้องดำเนินการตรวจสอบตัวอักษรถัดไปก่อน

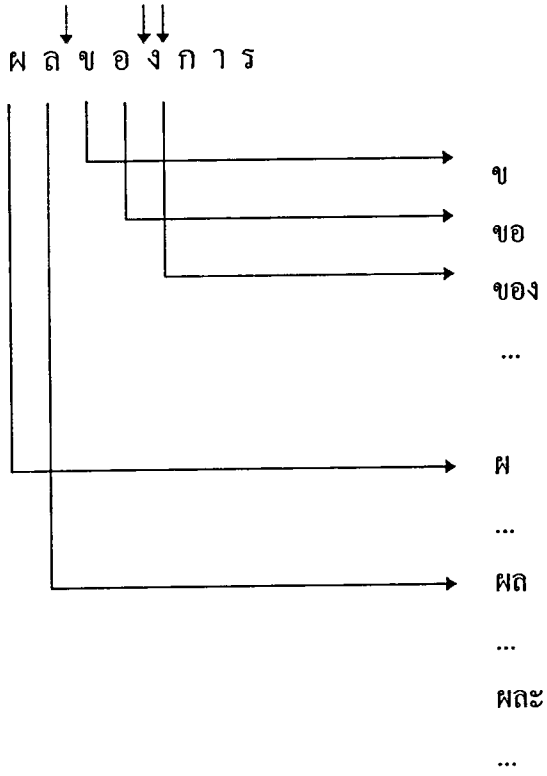
5. คำที่ไม่ถูกต้อง และไม่สามารถย้อนกลับได้ เช่น

นกมีเล็ป เมื่อตรวจสอบถึงตัวสุดท้าย คือ " ป " จะพบว่าเป็นคำที่ไม่มีในพจนานุกรม และไม่ได้เป็นส่วนหนึ่งของคำ เมื่อตรวจสอบการ mark ของคำก่อนหน้า คือ มี ก็ไม่พบการ mark จุดแยกคำ ดังนั้น จึงไม่สามารถย้อนกลับได้ ดังนั้น คำว่า เล็ป จึงเป็นคำผิด

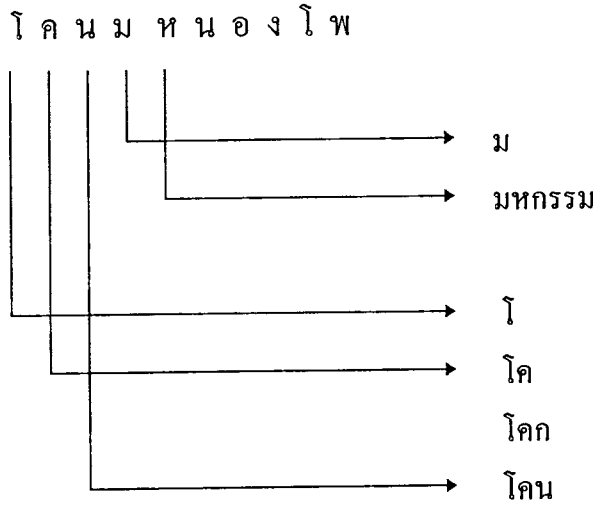
ผลงานวิจัยการตัดแบ่งคำไทยโดยการตรวจสอบกับพจนานุกรมที่ผ่านมา

1. ผลงานวิจัยของ อาจารย์ยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ เรื่อง " การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี " [8] ได้นำเสนอการตัดแบ่งคำ ดังนี้

ประโยค " ผลของการวิจัยนี้... "



ในการตัดคำจะเริ่มที่ " ผ " จะชี้ในหมวด " ผ " ของพจนานุกรมและนำตัวถัดไปคือ " ล " มาทำการเปรียบเทียบจะพบคำว่า " ผล " และเมื่อนำตัวอักษรถัดไปมาเปรียบเทียบคือ " ข " พบว่า " ผลข " ไม่พบในพจนานุกรม โปรแกรมจะทราบขอบเขตของคำทันที และจะนำตัวอักษร " ข " มาเริ่มเปรียบเทียบกับพจนานุกรมใหม่ พร้อมทั้งได้นำเสนอวิธีการตัดคำที่คลุมเครือด้วยวิธีการย้อนกลับ (Backtracking) ดังนี้ ประโยค " โคนมหนองโพ "



ในการตัดแบ่งคำพบว่า " โค " มีในพจนานุกรม และเมื่อหาคำต่อไปพบว่า " โคน " มีในพจนานุกรมทำให้มีโอกาสในการตัดคำและเป็นคำที่ยาวที่สุด และเมื่อค้นหาต่อพบว่า " โคนม " ไม่มีในพจนานุกรมจะทำการตัดคำ โคน / ม ก่อน(ซึ่งไม่ถูกต้อง) แต่เมื่อค้นหา " มหน " จะพบว่าไม่มีในพจนานุกรม ทำให้ควรมีการย้อนกลับไปตำแหน่งก่อนหน้าที่สามารถตัดคำได้ โดยถือว่า โค / นมหน แล้วทำการตรวจสอบ " นมหน " ต่อจะพบว่าตัดแบ่งได้เป็น นม / หน

2. ผลงานวิจัยของ ดร. รัตติกร วรากูลศิริพันธ์และสมศักดิ์ จันวัน เรื่อง " การตัดคำ จากประโยคในภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด " [12] โดยนำแนวทางของอาจารย์ยืน ภู่วรวรรณในการตัดคำที่ยาวที่สุดมาช่วยในการออกแบบ วิธีการตัดคำมีดังนี้ เช่น ประโยค " ฉันกินข้าว " ในการตัดแบ่งคำจะทำการหา ASCII ตัวแรกของประโยคคือ " ฉ " (169) และหา Weight ของประโยคโดยการนำค่า ASCII ของตัวอักษรแรกรวมกับตัวอักษรถัดไป คือ " ฉ " (169) + " ั " (209) = " ฉั " (378) แล้วทำการเปรียบเทียบโดยใช้ตัวอักษร 2 ตัวอักษรแรกของประโยคเพื่อชี้ไปในพจนานุกรมของคำที่ขึ้นต้นด้วยตัวอักษร 2 ตัวนั้น จากนั้นก็เริ่มเปรียบเทียบจากคำที่ยาวที่สุดก่อน คือ คำว่า " ฉันกินข้าว " ถ้าไม่พบในพจนานุกรมก็จะหาคำว่า " ฉันกินข้า " จนกว่าจะพบคำในพจนานุกรม ในที่นี้จะเท่ากับ 3 ตัวอักษร (ฉัน) แล้วทำการหาค่า ASCII 2 ตัวอักษรถัดไปแล้วทำการเปรียบเทียบกับพจนานุกรม จนสิ้นสุดประโยค

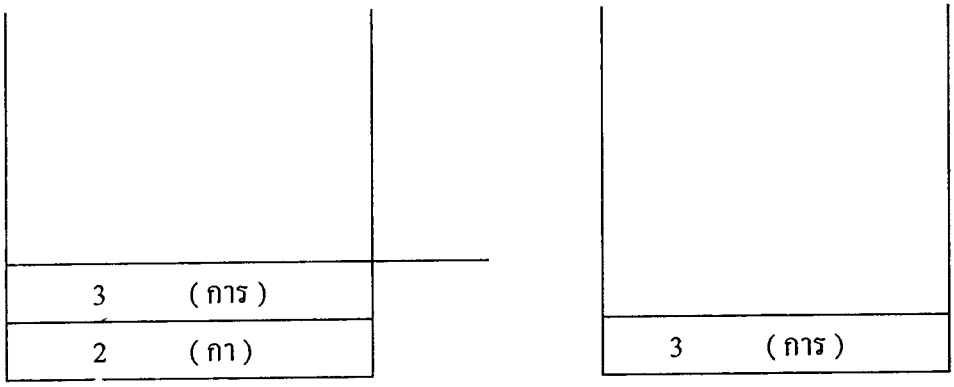
<u>หลักการตัดแบ่งคำภาษาไทยในวิทยานิพนธ์โดยใช้โครงสร้างพจนานุกรมแบบแฮชชิ่ง</u>

ในการตัดแบ่งพยางค์จากพจนานุกรมนี้จะใช้หลักการเปรียบเทียบคำจากพจนานุกรมที่มีอยู่ในหน่วยความจำ โดยในวิทยานิพนธ์ฉบับนี้ ได้นำแนวทางการตัดคำของอาจารย์ยืน ภู่วรวรรณตามหลักการแบบ Backtracking มาทำการออกแบบและนำเสนอวิธีการตัดคำโดยใช้โครงสร้างในการตัดคำอีกแบบหนึ่ง ซึ่งจะได้นำเสนอโครงสร้างข้อมูลแบบ Stack เข้ามาช่วยในการตัดแบ่งคำภาษาไทยด้วย ซึ่งจะทำให้มีการทำงานของการตัดแบ่งคำเป็นแบบ Iterative ซึ่งแตกต่างกับการตัดคำที่เป็นแบบ Recursive เดิม วิธีการนี้จะนำตัวอักษรมาทำการคำนวณผ่าน Hashing Function (เดียวกับโครงสร้างพจนานุกรมที่ได้ออกแบบไว้ในบทที่ 3) ทีละตัวอักษร แล้วนำมาเปรียบเทียบกับคำในโครงสร้างพจนานุกรม แบบแฮชชิ่ง ที่ได้ออกแบบไว้ว่าพบหรือไม่พบคำ ๆ นั้น ถ้าพบจะเก็บตำแหน่งของตัวอักษรนั้นไว้ใน Stack ตัวที่ 1 และทำการค้นหาต่อจนกว่าจะพบคำที่ยาวที่สุดในการตัดแบ่งคำได้ในแต่ละรอบ ซึ่งจะนำตำแหน่งของตัวอักษรที่ยาวที่สุดในแต่ละรอบมาเก็บไว้ใน Stack ตัวที่ 2 แล้วนำตัวอักษรถัดไปต่อจากตำแหน่งที่ยาวที่สุดมาทำการคำนวณต่อ

จนกว่าจะสิ้นสุดของประโยคหรือสิ้นสุดบรรทัดของประโยคนั้น ๆ หรือพบว่าคำที่กำลังเปรียบเทียบในพจนานุกรมนั้นไม่มีตัวต่อตามโครงสร้างพจนานุกรมภาพที่ 3.1 (word.continue) จากหลักการตัดแบ่งคำนี้จะขอยกตัวอย่างเพื่อให้รายละเอียดได้ชัดเจนขึ้น

เช่น	ตำแหน่ง	1	2	3	4	5	6	7	8	9	10
	ตัวอักษร	ก	า	ร	ต	ร	ว	จ	ส	อ	บ

รอบที่ 1 จะนำตัวอักษร "ก" มาทำการคำนวณผ่าน Hashing Function ก่อน เมื่อเปรียบเทียบคำกับโครงสร้างพจนานุกรม ถ้าไม่พบจะนำตัวอักษรถัดไป คือ "า" มาทำการคำนวณต่อ และเปรียบเทียบ ซึ่งพบว่าคำว่า "กา" มีในพจนานุกรมทำการในตำแหน่งใน Stack ตัวที่ 1 ดังภาพที่ 4.1 และทำการนำตัวอักษรถัดไปคือ "ร" มาทำการคำนวณต่อ พบว่า คำว่า "การ" มีในพจนานุกรม ใส่ตำแหน่งใน Stack ตัวที่ 1 และนำตัวอักษรถัดไปมาคำนวณต่อ จนจบประโยค พบว่า "การ" เป็นคำที่ยาวที่สุดในรอบที่ 1 ใส่คำตำแหน่งที่ยาวที่สุดใน Stack ตัวที่ 2 ดังภาพที่ 4.1

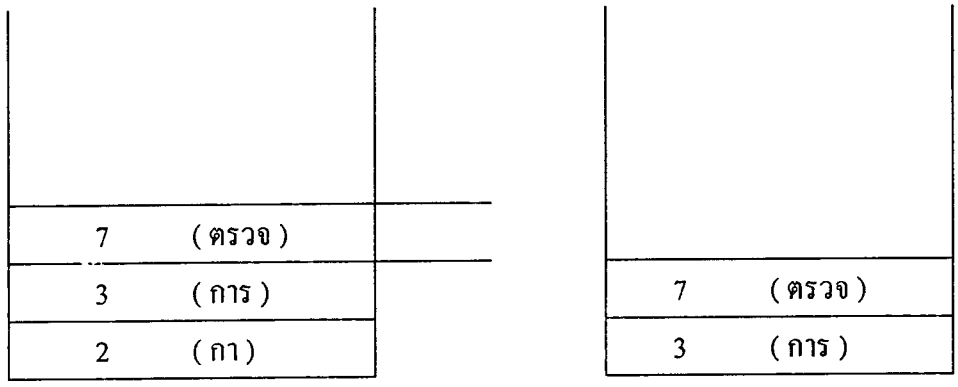


Stack ตัวที่ 1
ตำแหน่งที่ตัดได้

Stack ตัวที่ 2
ตำแหน่งที่ยาวที่สุด

ภาพที่ 4.1 แสดงค่าตำแหน่งที่สามารถทำการตัดได้

รอบที่ 2 นำตัวอักษรที่อยู่ในตำแหน่งที่ยาวที่สุด มาทำการคำนวณต่อ โดยตัวอักษรที่นำมาคือ " ต " แล้วเปรียบเทียบคำในพจนานุกรม และนำตัวอักษรถัดไป ทำคำนวณต่อไปเรื่อย ๆ ซึ่งจะพบว่าคำที่ยาวที่สุดในรอบนี้ คือ " ตรวจ " ซึ่งไม่มีตำแหน่งที่ตัดได้ก่อนหน้า จึงนำค่ามาใส่ Stack ทั้ง 2 ตัว ดังภาพที่ 4.2

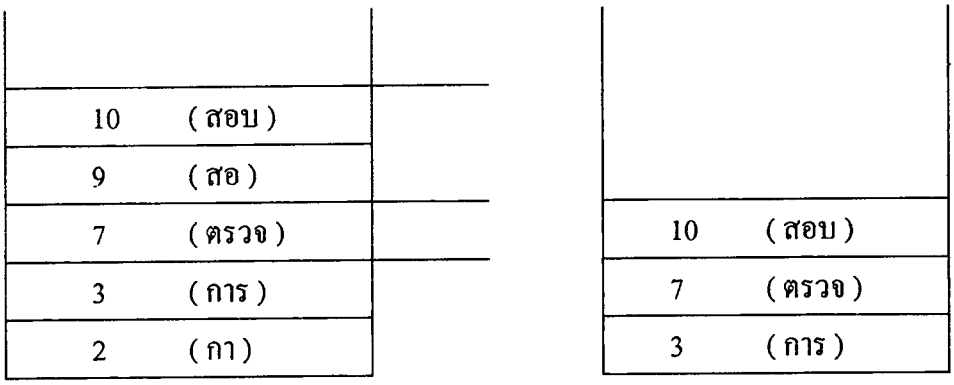


Stack ตัวที่ 1
ตำแหน่งที่ตัดได้

Stack ตัวที่ 2
ตำแหน่งที่ยาวที่สุด

ภาพที่ 4.2 แสดงค่าตำแหน่งที่สามารถตัดคำได้

รอบที่ 3 นำตัวอักษรถัดไปในรอบที่ 2 มาคำนวณต่อ โดยนำ " ส " มาคำนวณเปรียบเทียบคำในพจนานุกรม ซึ่งจะพบว่า มีคำว่า " สอ " สามารถตัดคำได้ และคำว่า " สอบ " ตัดคำได้ ก็จะนำค่าตำแหน่งไปใส่ใน Stack ตัวที่ 1 และค่าที่ยาวที่สุดใส่ใน Stack ตัวที่ 2 ดังภาพที่ 4.3



ภาพที่ 4.3 แสดงค่าตำแหน่งที่สามารถตัดคำได้

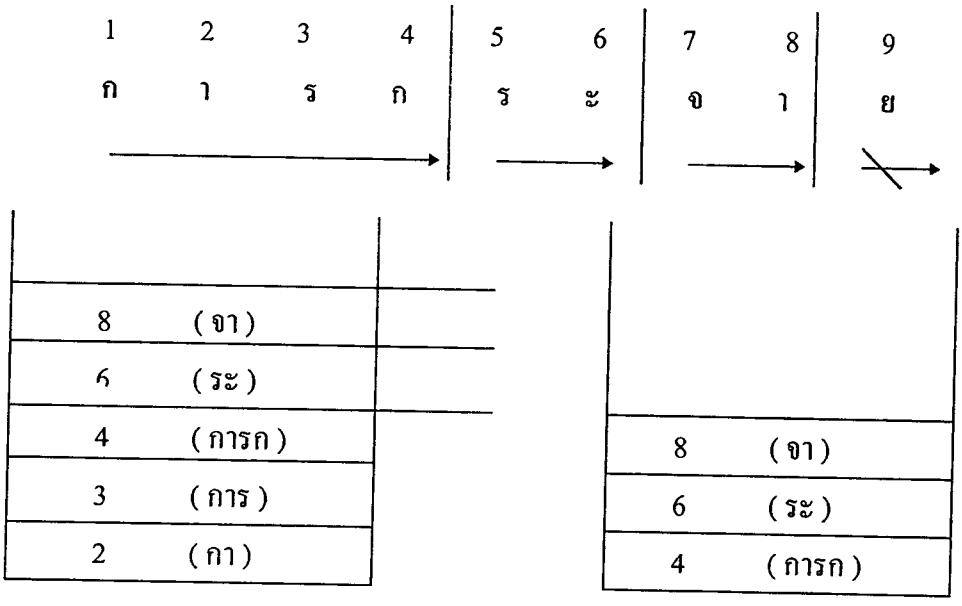
<u>การถอยหลังเมื่อพบปัญหาหรือคำที่คลุมเครือ (Back Tracking)</u>

ในการค้นหาคำในพจนานุกรมเพื่อนำมาตัดแบ่งคำ อาจจะพบปัญหาในการตัดแบ่งคำบางคำ ซึ่งในงานวิจัยนี้ได้นำวิธีการ Back Tracking มาใช้เพื่อใช้ ตัดสินการตัดแบ่งคำ เพื่อให้ได้คำที่ถูกต้องมากที่สุด ดังตัวอย่าง

ตำแหน่ง	1	2	3	4	5	6	7	8	9
ตัวอักษร	ก	า	ร	ก	ร	ะ	จ	า	ย

ในการทำงานจะมีขั้นตอนดังนี้ คือ เริ่มต้นจะหาคำที่ยาวที่สุดก่อน ในที่นี้ จะได้คำว่า "การก" และทำการเก็บในแต่ละตัวใน Stack ตัวที่ 1 ว่ามีคำใดที่สั้นกว่าคำว่า "การก" และพบในพจนานุกรม คือคำว่า กา, การ จากนั้นในรอบต่อไปจะนำตัวอักษรถัดจากคำว่า การก มาตัดแบ่งคำต่อ จะได้คำว่า "ระ" และนำตัวอักษรถัดไปมาตัดแบ่งต่อ จะได้คำว่า "จา" แล้วนำตัวอักษรถัดไปมาตัดแบ่ง ในที่นี้จะเหลือ "ย" ตัวเดียว ซึ่งไม่สามารถตัดแบ่งได้ จึงต้องทำการ Back Track ไปยังตำแหน่งก่อนหน้า จนกว่าจะทำการ Back track หมดตำแหน่งก่อนหน้าใน Stack ตัวที่ 1 หรือ Back track จนกว่าจะพบคำที่ตัดได้ในประโยคต่อ โดยจะมีขั้นตอนการทำงานดังภาพที่ 4.4

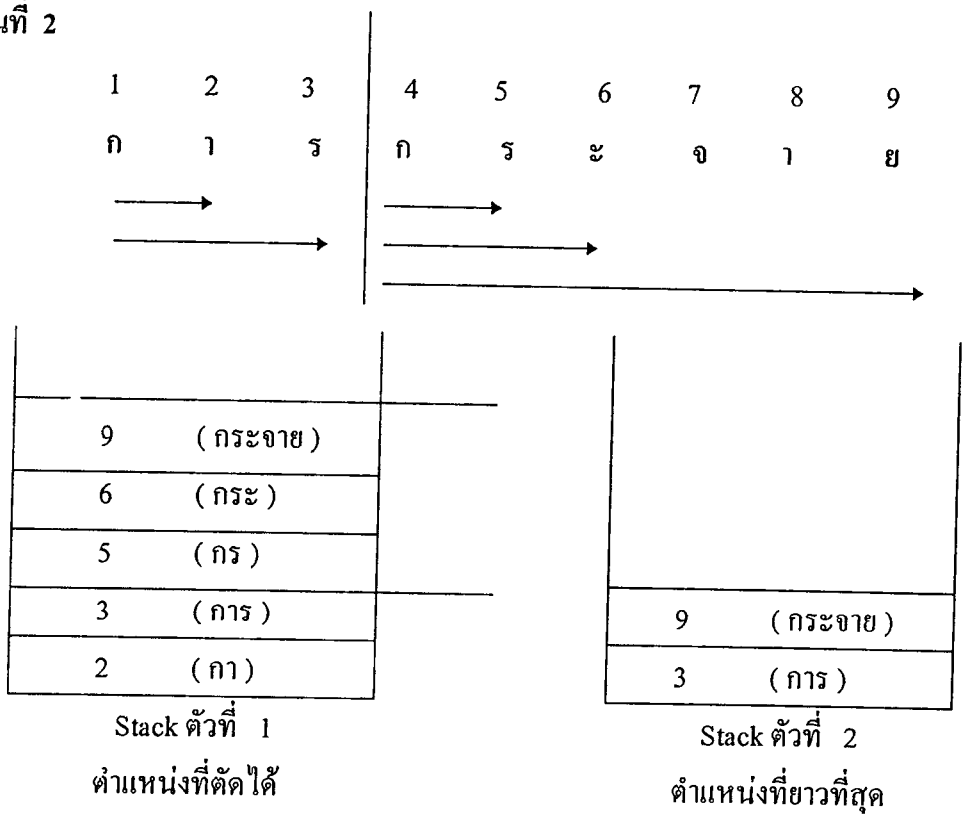
ขั้นตอนที่ 1



Stack ตัวที่ 1
ตำแหน่งที่ตัดได้

Stack ตัวที่ 2
ตำแหน่งที่ยาวที่สุด

ขั้นตอนที่ 2



Stack ตัวที่ 1
ตำแหน่งที่ตัดได้

Stack ตัวที่ 2
ตำแหน่งที่ยาวที่สุด

ภาพที่ 4.4 แสดงขั้นตอนตำแหน่งการตัดคำใน Stack ในการ Back Tracking

<u>อัลกอริทึมการตัดแบ่งคำภาษาไทย</u>

```
pos = 0 ;  count  =  0 ;
While ( text [ pos ] ! = ตัวหยุดประโยค )          -----  จะตรวจสอบการสิ้นสุดประโยค
  { For ( h = maxposition = 0;q  =  pos;q  <=  count ;q+ + )   -----  loop ตรวจสอบคำหยุด
                                                          เมื่อ count = 25 หรือไม่พบตัวต่อ
      { p  =  text  [ q ] ;                        -----  p จะเป็นตัวชี้ตำแหน่งตัวอักษร
        hash  =  Horner  ( h, p ) ;                -----  คำนวณการ hashing
        word  = Dict (hash, text, pos, q, found) ;        -----  จะส่งตำแหน่งที่สามารถ
                                                          ตัดได้ ถ้าสามารถตัดได้ จะให้
                                                          found = 1 เพื่อจะเก็บตำแหน่งที่
                                                          ยาวที่สุดต่อไป
        h = hash;                                  -----  ส่งค่ากลับมาคำนวณต่อ
        If (found)
          {  Pushstack1  ( q );                    -----  เก็บตำแหน่งที่ตัดได้ใน Stack ตัวที่ 1
             maxposistion = q;                     -----  เก็บตำแหน่งที่ยาวที่สุด
             If ( word.continue == -1)             -----  ตรวจสอบตัวอักษรที่มีต่อจากคำ
                  { break; }
          }                                        -----  If Loop
      }                                            -----  For Loop
    If ( maxposition  !=  0)                       -----  พบตำแหน่งที่ยาวที่สุด
      { Pushstack2  (maxposition)                  -----  เก็บตำแหน่งที่ยาวที่สุดในStack ตัวที่ 2
        pos  =  maxposition ;                      -----  ให้ตำแหน่งที่จะคำนวณต่อหลังจากที่
                                                          ตัดคำได้
      }
     Else  { pos =  Popstack ( ) ; }  .            ------  ให้ตำแหน่ง Back Track
                                                          กลับไปก่อนหน้าเมื่อพบคำกำกวม
  }      //   จบ   while  loop
```

หมายเหตุ ตัวหยุดประโยค คือ ตัวที่ขึ้นประโยคใหม่ (\n) หรือ ตัวที่เป็น Space (' ')

โดยที่ pos , q ----- ตำแหน่งของตัวอักษรในประโยค

count ----- จำนวนตัวอักษรที่จะตรวจสอบในประโยค (25 ตัวอักษร)

text ----- array เก็บตัวอักษรในประโยค

p ----- ตัวชี้ตัวอักษร

hash ----- ค่าแอดเดรสของคำ

found ----- ตรวจสอบว่าพบคำที่สามารถตัดได้ในประโยค

maxposition ----- ตำแหน่งของตัวอักษรที่ยาวที่สุดที่ตัดได้

h ----- ค่า hashing เพื่อส่งค่ากลับไปคำนวณต่อ

word.continue ----- สำหรับตรวจสอบตัวอักษรที่ต่อจากคำที่เปรียบเทียบ

Horner ----- Hashing Function

Dict ----- เป็น function เปรียบเทียบคำในพจนานุกรม

Pushstack1 ----- เป็น function ในการใส่ตำแหน่งที่สามารถตัดได้

Pushstack2 ----- เป็น function ในการใส่ตำแหน่งที่ยาวที่สุดที่สามารถตัดได้

PopStack ----- เป็น function Backtrack เพื่อจะได้คำที่ถูกต้องมากที่สุด

หมายเหตุ ในการ PopStack จะมีการตรวจคำที่ตัดก่อนหน้าว่ามีการตัดมากกว่า 1 ครั้งในแต่ละรอบ (ตรวจสอบใน Stack ตัวที่ 1) หรือไม่ ถ้าไม่จะทำการ Popstack ตำแหน่งใน Stack ตัวที่ 1 ไปจนกว่าจะพบการตัดตำแหน่งของคำก่อนหน้ามากกว่า 1 ครั้ง เพื่อนำตำแหน่งนั้นมาทำการเปรียบเทียบกับพจนานุกรมต่อไปโดยการตัดคำรอบใหม่จะทำการตรวจสอบตำแหน่งที่ตัดได้กับตำแหน่งใน Stack ตัวที่ 2 ถ้าพบว่าสามารถตัดคำได้มากกว่าในรอบก่อนจะทำการเปลี่ยนตำแหน่งใน Stack ตัวที่ 2 เพื่อจะได้คำที่ยาวที่สุดในการตัดคำ

จากอัลกอริทึมการตัดคำจะต้องทำการค้นหาคำในพจนานุกรมก่อน ดังนั้นค่าเฉลี่ยในการค้นหาจะเท่ากับผลคูณของค่าเฉลี่ยในการค้นหาที่เกิด Collision (ตารางที่ 3.1) กับค่าเฉลี่ยของจำนวนตัวอักษรที่ประกอบเป็นคำ (ตารางที่ 3.2) ซึ่งจะได้ค่าเป็น $1*n + 0.5*6.76$ โดย n เป็นจำนวนตัวอักษรที่เปรียบเทียบ ซึ่งค่ามากที่สุดในการค้นหา (worst case) เท่ากับ $1*n + 5*6.76$ และค่าที่น้อยที่สุดในการค้นหา (best case) เท่ากับ $1*n$

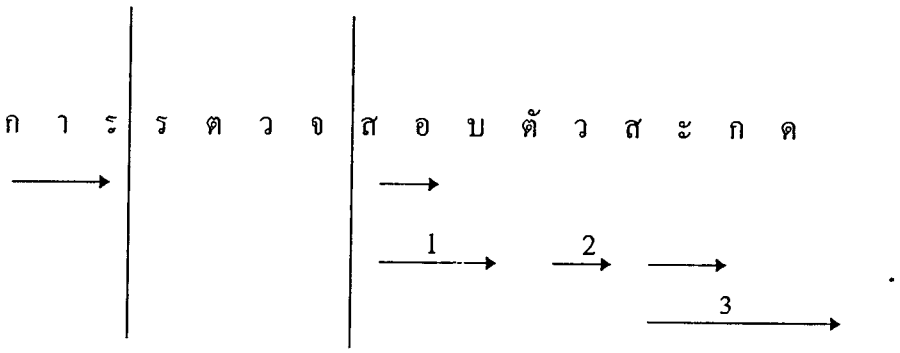
<u>การตัดแบ่งคำภาษาไทยเมื่อพบคำผิดในเอกสาร</u>

ลักษณะของคำที่เกิดในงานเอกสาร [20], [21]

1. การพิมพ์เกิน เช่น การ<u>ตตรวจ</u>สอบตัวสะกด
2. การพิมพ์ผิด เช่น การ<u>ดรวจ</u>สอบตัวสะกด
3. การพิมพ์ขาด เช่น การ<u>รวจ</u>สอบตัวสะกด
4. การพิมพ์ที่มีการสลับที่ระหว่างตัวอักษรที่ถูกในคำ เช่น การ<u>รตวจ</u>สอบตัวสะกด

<u>การหาขอบเขตของคำที่ไม่พบในพจนานุกรม</u>

ในการตัดแบ่งคำภาษาไทยเปรียบเทียบกับโครงสร้างพจนานุกรม เมื่อทำการ BackTrack จนไปถึงที่ตัวอักษรต้นบรรทัดก่อนหน้า โปรแกรมจะทำการหาขอบเขตของคำที่ไม่พบในพจนานุกรมโดยการเก็บตำแหน่งสุดท้ายที่สามารถตัดได้มากที่สุดเป็นตำแหน่งเริ่มต้นของคำที่ผิด แล้วทำการ Skip ตัวอักษรไปทีละตัว เพื่อนำตัวอักษรถัดไปมาทำการคำนวณ เปรียบเทียบกับพจนานุกรม จนกว่าตำแหน่งที่คำนวณใหม่สามารถทำการตัดแบ่งคำได้มากกว่าหรือเท่ากับ สองคำที่พบในพจนานุกรม โปรแกรมจะนำตำแหน่งที่เริ่มคำนวณนั้นมาเป็นจุดสิ้นสุดคำที่ผิด เพื่อทำการค้นหาคำคล้ายในการแก้ไขต่อไป โดยนำคำผิดมาทำการค้นหาโครงสร้างพจนานุกรมคำคล้ายที่ได้ออกแบบไว้ ในบทที่ 3 ดังตัวอย่างต่อไปนี้



จากตัวอย่างจะพบว่า " รตวจ " เป็นคำผิด จะนำตัวไปหาคำคล้ายในโครงสร้างพจนานุกรมคำคล้าย ดังรูปที่ 3.6 ในบทที่ 3 จะพบคำว่า รตนะ , รตะ , ตรวจ , ตำรวจ , สำรวจ เป็นต้น เพื่อให้ผู้ใช้เลือกทำการแก้ไขต่อไป

บทที่ 5

ผลการทดสอบการตัดแบ่งคำและเปรียบเทียบกับ Microsoft Word

ในการวิเคราะห์การทำงานของโปรแกรมการตัดคำภาษาไทยโดยใช้พจนานุกรมที่มีโครงสร้างแฮชชิ่งในวิทยานิพนธ์ จะพบว่ามีความสามารถในการตัดแบ่งคำได้อย่างถูกต้องและรวดเร็วกว่าโปรแกรม Commerial Microsoft Word

<u>ข้อมูลที่ใช้ในการทดสอบ</u>

คำศัพท์ที่ใช้สำหรับการสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย เป็นคำศัพท์พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 ซึ่งมีรายการคำทั้งหมด 32,381 คำนำมาสร้างพจนานุกรมโดยมีการสร้าง Index ในรายการคำ เพื่อใช้ในการตรวจสอบการตัดคำภาษาไทย ซึ่งผู้วิจัยได้พบว่ารายการคำในพจนานุกรมของ Microsoft Word (ดังตารางรายการคำในภาคผนวก ข) มีจำนวนรายการคำน้อยกว่า รายการคำในวิทยานิพนธ์ โดยผู้วิจัยได้ทำการทดสอบการตัดคำในประโยคเดียวกัน ที่มีรายการคำเหมือนกันทั้งในงานวิจัยและ Microsoft Word ก็จะพบข้อผิดพลาดของ Microsoft Word ในการตัดคำ และในการตัดคำของ Microsoft Word ยังพบคำที่ถูกต้องตามพจนานุกรม แต่ Microsoft Word ตรวจสอบเป็นคำผิด เช่น คำว่า อุตสาหกรรม, แม่นยำ, เศรษฐกิจ เป็นต้น ดังผลการทดสอบในตารางที่ 5.1

ตารางที่ 5.1 แสดงจำนวนคำจากพจนานุกรมราชบัณฑิตยสถานที่ไม่พบในพจนานุกรมของ Microsoft Word

โปรแกรม	จำนวนคำทั้งหมด	คำที่ตรวจสอบ			
		ถูก	% ที่ ถูก	ผิด	% ที่ผิด
Word 6.0	32,381	30,090	92.92	2,291	7.08
Word 7.0	32,381	31,374	96.89	1,007	3.11

<u>การทดสอบการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทยกับ Microsoft Word</u>

การทดสอบการตัดคำในงานวิจัยกับ Microsoft Word จะสามารถแบ่งการเปรียบเทียบได้ดังนี้

ก) ความเร็วในการตัดคำ

ข) ความถูกต้องในการตัดคำ

ค) การค้นหาคำคล้ายเมื่อพบคำผิดในเอกสาร

ก) ความเร็วในการตัดคำ

จากการที่ได้ทำการวัดความเร็วในการตรวจสอบเอกสารชนิดเดียวกัน(เอกสารที่ตรวจสอบคือบทคัดย่อภาษาไทยในวิทยานิพนธ์ฉบับนี้ ซึ่งทำการคัดลอกจนได้จำนวนหน้าตามที่ต้องการทดสอบ) ระหว่าง Microsoft Word กับโปรแกรมของงานวิจัยจะได้ผลดังตารางที่ 5.2

ตารางที่ 5.2 ตารางแสดงความเร็วในการตรวจสอบเป็นวินาที (บน Pentuim 90 MHz)

จำนวนหน้า / จำนวนบรรทัด / Software	5 145	10 288	20 575	50 1436	100 2868	200 5736
Word 6.0	3	4	9	20	38	75
โปรแกรม Spell Check 1	2	3	6	14	28	55
โปรแกรม Spell Check 2	1	2	4	7	13	26
โปรแกรม Spell Check 3	1	2	4	9	17	35
โปรแกรม Spell Check 4	0.90	1.50	2.72	6.17	11.82	23.20
โปรแกรม Spell Check 5	0.81	1.28	2.28	5.25	10.14	19.92

จากตารางที่ 5.2

โปรแกรม Spell Check 1 - พิจารณาตรวจสอบจนสิ้นสุดประโยค

โปรแกรม Spell Check 2 - พิจารณาตรวจสอบทีละ 15 ตัวอักษรของประโยค

โปรแกรม Spell Check 3 - พิจารณาตรวจสอบทีละ 25 ตัวอักษรของประโยค

โปรแกรม Spell Check 4 - ทำการเปลี่ยนโครงสร้างในการ PopStack โดยพิจารณาจำนวนรอบที่สามารถตัดได้ในแต่ละรอบ ถ้าในรอบมีตำแหน่งที่สามารถตัดได้เพียง 1 ครั้ง ให้ทำการ PopStack ต่อไปอีก

โปรแกรม Spell Check 5 - โครงสร้างพจนานุกรมเพิ่มฟิลด์แสดงตัวอักษรที่มีตัวต่อของคำที่พิจารณา(ดังภาพที่ 3.1)และพิจารณาตรวจสอบไป 25 ตัวอักษรของประโยค

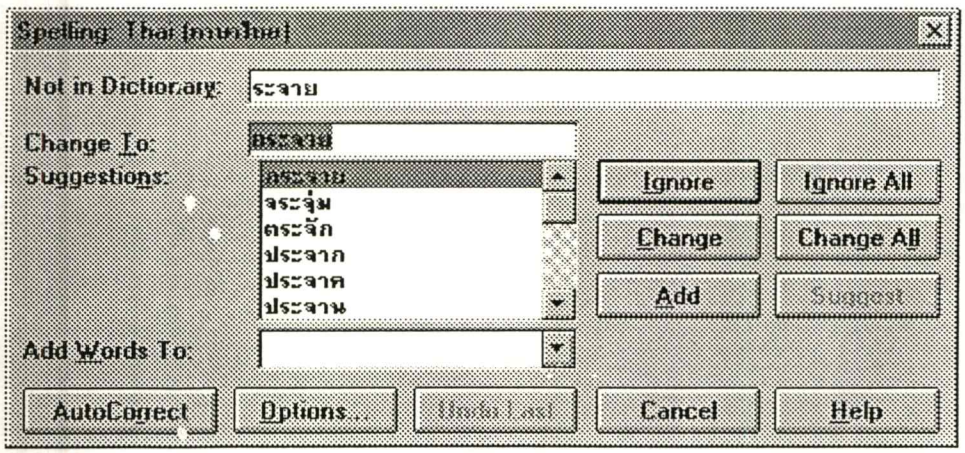
หมายเหตุ การทดสอบกับ Microsoft Word 7.0 ไม่สามารถตรวจสอบได้เพราะโปรแกรมไม่มีการตอบสนองในการ Spell Checking

ข) ความถูกต้องในการตัดคำ

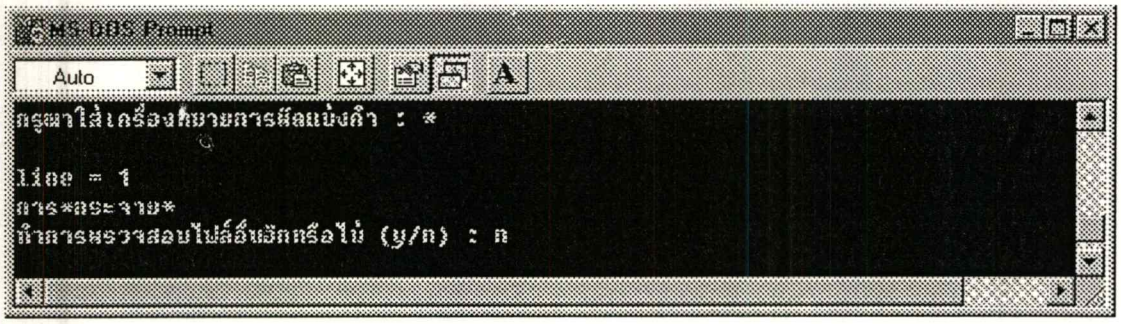
Input การกระจาย

Output

Word การก**ระจาย**



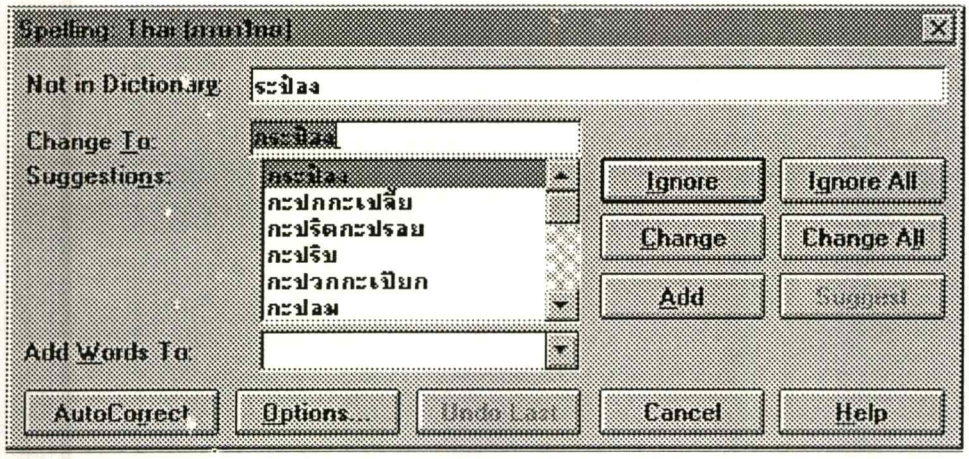
งานวิจัย



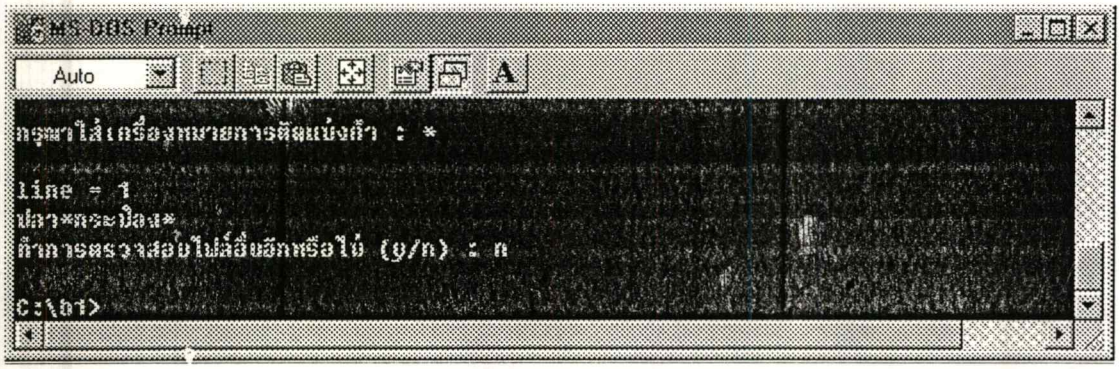
Input ปลากระป๋อง

Output

Word ปลาก**ระป๋อง**



งานวิจัย



ตัวอย่างประโยคที่ Microsoft Word ตัดแบ่งคำผิด

ตัวอย่างประโยค	Word 6.0	Word 7.0
ผู้นำเข้าจะเข้มงวดมาก	ผู้นำเข้าจะเข้ มงวด มาก	ผู้นำเข้าจะเข้ มงวด มาก
ปลาทูนากระป๋อง	ปลาทูนาก ระป๋อง	ปลาทูนาก ระป๋อง
ขอบริจาคหนังสือ	ขอบ ริจาค หนังสือ	ขอบ ริจาค หนังสือ
การกระจาย	การก ระจาย	การก ระจาย
มาบริจาค	มาบ ริจาค	มาบ ริจาค
ปลากระป๋อง	ปลาก ระป๋อง	ปลาก ระป๋อง
จะมองหาสมาชิก	จะมองหาส มาชิก	จะมองหาส มาชิก
จะจ้างงาน	จะจ้า งงาน	จะจ้า งงาน

จากตัวอย่างประโยคที่แสดงให้เห็นการตัดคำระหว่าง Microsoft Word กับงานวิจัย จะพบว่าการทำงานของ Microsoft Word จะตัดคำผิดซึ่งคำในประโยคเหล่านั้นเป็นคำที่ถูกต้องตามพจนานุกรม โดยถ้านำคำจากประโยคข้างต้นมาทำการแยกเป็นคำๆจะพบว่าในส่วนของ Microsoft Word ก็จะสามารถตัดแบ่งคำได้ถูกต้อง

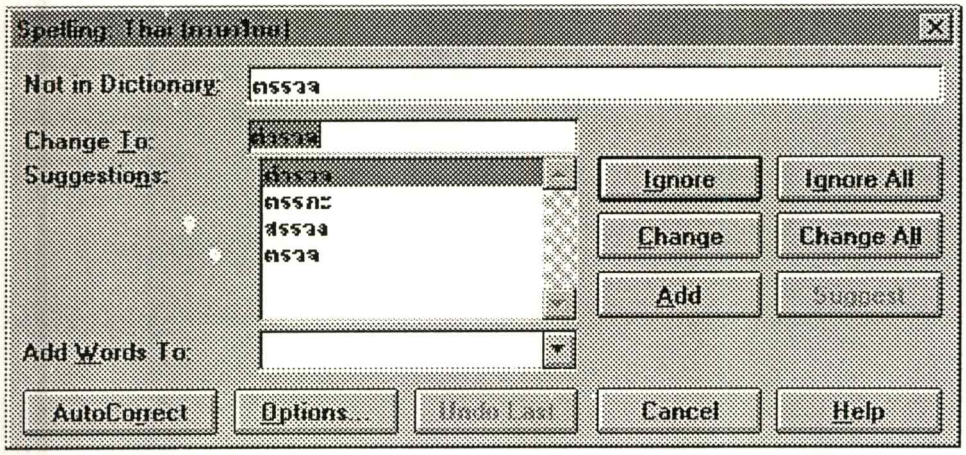
ค) การค้นหาคำคล้ายเมื่อพบคำผิดในเอกสาร

ในการค้นหาคำคล้ายในงานวิจัยจะหาจากการนำเอาขอบเขตคำที่ผิดมาหาในพจนานุกรมโดยทำการแบ่งครึ่งความยาวของขอบเขตคำผิด จะหาคำคล้ายโดยการนำคำที่แบ่งครึ่งของคำที่ผิดมาทำการค้นหาทั้งสองด้านของคำ (รายละเอียดในบทที่ 3) แต่ใน Microsoft Word จะใช้วิธีการค้นหาโดยการนำคำที่ผิดมาพิจารณาตัวอักษรเป็นคู่ แล้วทำการค้นหาคำในพจนานุกรม

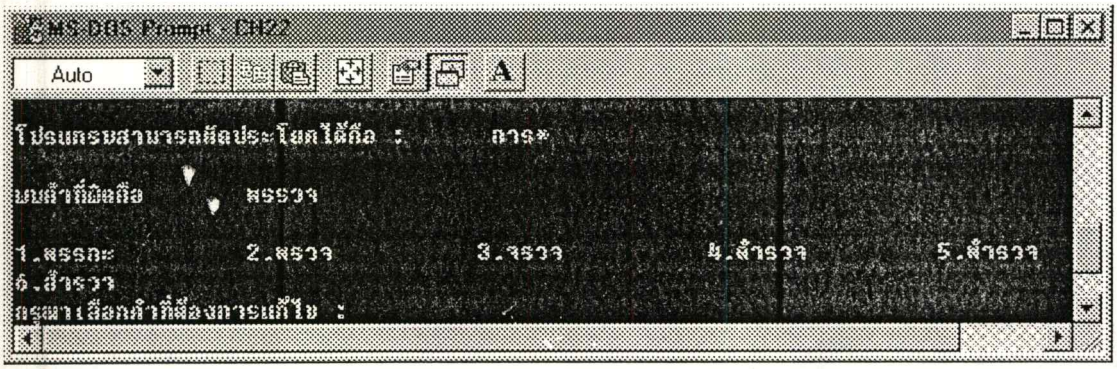
ตัวอย่างการค้นหาคำคล้าย

ประโยค " การตรรวจสอบ "

Word



งานวิจัย



จากการทดลอบจะพบว่าการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ฉบับราชบัณฑิตยสถาน จะมีความถูกต้องสูง สามารถทำการตรวจสอบได้รวดเร็ว และยังมีการค้นหาคำคล้ายได้ใกล้เคียงกับคำที่ผิดในเอกสารมาก ซึ่งเป็นข้อที่ได้เปรียบของการตัดคำภาษาไทยเมื่อทำการเปรียบเทียบกับโปรแกรม Microsoft Word ที่ได้นำเสนอในวิทยานิพนธ์นี้

บทที่ 6

บทสรุปและข้อเสนอแนะ

บทสรุป

ในการวิจัยทางด้านภาษาศาสตร์ ได้ มีนักวิจัยหลายท่านให้ความสนใจและมีการพัฒนาต่อเนื่องกันเรื่อยๆมา โดยการตัดแบ่งคำภาษาไทยนี้เป็นส่วนหนึ่งของงานทางด้านภาษาศาสตร์ ซึ่งในวิทยานิพนธ์นี้ได้นำเสนออัลกอริทึมในการสืบค้นคำศัพท์ที่มีโครงสร้างข้อมูลแบบแฮชชิ่ง (Hashing) มาทำการเก็บคำในพจนานุกรมและยังใช้อัลกอริทึมแฮชชิ่งนี้มาทำการตัดแบ่งคำภาษาไทย โดยความสามารถในการตัดคำจะให้ผลลัพธ์เป็นคำศัพท์ที่ยาวที่สุดที่พบในพจนานุกรม และสามารถตัดคำได้ทุกคำที่ปรากฏเป็นคำศัพท์ในพจนานุกรม ในด้านความเร็วในการตรวจสอบเอกสารจะมีความเร็วกว่าโปรแกรม Microsoft Word

โครงสร้างข้อมูลสำหรับการเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่พัฒนาขึ้นมาโดยใช้โครงสร้างข้อมูลแบบแฮชชิ่งนี้ จะมีประสิทธิภาพที่ดีในด้านของการเพิ่มเติมคำศัพท์ใหม่ สามารถสร้าง Index ของคำได้ง่าย และยังสามารถค้นหาคำคล้ายได้สะดวกรวดเร็ว โดยในการตัดคำในวิทยานิพนธ์นี้ได้ทำการออกแบบอัลกอริทึมที่เป็นแบบลักษณะ Iterative และใช้โครงสร้างข้อมูลแบบ Stack ในการเก็บตำแหน่งคำที่ตัดได้แทนอัลกอริทึมแบบ Recursive ซึ่งเป็นแบบเดิมที่ได้มีผู้ทำการวิจัยมาแล้วซึ่งจะทำให้การตัดคำจะมีการตรวจสอบคำได้ทุกตำแหน่งก่อนทำการตัดคำ

ข้อเสนอแนะ

วิทยานิพนธ์นี้ได้ทำการสร้างพจนานุกรมอิเล็กทรอนิกส์ที่บรรจุคำ 32,381 คำ ในทางปฏิบัติการกับคำในพจนานุกรม จะต้องเก็บคำทั้งหมดในหน่วยความจำ ถ้าจำนวนคำที่จะเกิดขึ้นต่อไปมีจำนวนมากขึ้น จะทำให้หน่วยความจำที่มีอยู่จำกัดไม่สามารถรองรับข้อมูลทั้งหมดได้ แต่ถ้ามีการปรับปรุงการจัดการหน่วยความจำกับจำนวนคำในพจนานุกรม ให้ใช้งานได้อย่างมีประสิทธิภาพโดยไม่ทำให้ประสิทธิภาพทางด้านอื่นๆลดลง จะทำให้การตรวจสอบการตัดคำภาษาไทยกับพจนานุกรมทำงานได้ดีมากขึ้น

สำหรับอัลกอริทึมที่ใช้ในการตัดแบ่งคำที่พัฒนาในวิทยานิพนธ์นี้ จะใช้อัลกอริทึมที่เปรียบเทียบกับพจนานุกรมเท่านั้น ถ้าต้องการให้อัลกอริทึมในการตัดคำมีประสิทธิภาพสมบูรณ์มากขึ้นควรจะนำกฎในการตรวจสอบคำในภาษาไทยมาผสมผสานกับการเปรียบเทียบคำในพจนานุกรม จะทำให้มีความถูกต้องในการตัดแบ่งคำมากขึ้น และจะทำให้งานทางด้านภาษาศาสตร์มีประโยชน์มากขึ้นในการนำไปใช้ต่อไป

บรรณานุกรม

1. " พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 " จาก CD-ROM พจนานุกรมฉบับราช-บัณฑิตยสถาน, ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC).
2. ทวีศักดิ์ จันทร์วิทยานุชิต " การตรวจสอบคำผิดในเอกสารภาษาไทย " โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, เล่มที่ 2, ปีงบประมาณ 2533.
3. ประกาศิต ชาติบุรุษ และอาทิตย์ จิตต์จุฬานนท์ โครงสร้างข้อมูลและอัลกอริทึม สำนักพิมพ์ ซีเอ็ด, 2521.
4. ปรีชา ทิชินพงศ์ ลักษณะภาษาไทย สำนักพิมพ์โอเดียนสโตร์, 2528.
5. มนตรี พจนารถลาวัณย์ การเขียนโปรแกรมคอมพิวเตอร์ด้วยเทอร์โบซี สำนักพิมพ์ ซีเอ็ด, 2521.
6. ยืน ภู่วรวรรณ " เทคนิคพัฒนาการภาษาไทยบนเครื่องไมโครคอมพิวเตอร์ " วารสารไมโครคอมพิวเตอร์, ฉบับที่ 3, 2528-2529.
7. ยืน ภู่วรวรรณ และชัยยงศ์ วงศ์ชัยสุวัฒน์ การประมวลผลภาษาธรรมชาติ (Natural Language Processing) สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี.
8. ยืน ภู่วรวรรณ และวิวรรธน์ อิ่มอารมณ์ " การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี " รายงานการประชุมทางวิชาการ วิศวกรรมไฟฟ้า สถาบันอุดมศึกษาแห่งประเทศไทย, ครั้งที่ 9 , ธันวาคม 2529.
9. ยืน ภู่วรวรรณ และวิวรรธน์ อิ่มอารมณ์ " การตรวจสอบตัวสะกดด้วยคอมพิวเตอร์ " รายงานการประชุมทางวิชาการวิศวกรรมไฟฟ้า สถาบันอุดมศึกษาแห่งประเทศไทย, ครั้งที่ 10 , พฤศจิกายน 2530.
10. ยืน ภู่วรวรรณ, สมนึก คีรีโตและ สุรศักดิ์ สงวนพงษ์ " การตัดพยางค์คำไทยด้วยโครงสร้างข้อมูลแบบต้นไม้ ", รายงานการประชุมวิชาการ ทางวิศวกรรมไฟฟ้า 8 สถาบันอุดมศึกษา ครั้งที่ 8, ธันวาคม 2528.
11. ยืน ภู่วรวรรณ และเสรี เศวตเศรนี " การพัฒนาอัลกอริทึม สำหรับประมวลผลตัวอักษรไทย ", วิศวกรรมสาร ปีที่ 37, เล่มที่ 3, มิถุนายน 2527.

12. รัตติกร วราภูลศิริพันธ์และ สมศักดิ์ จันวัน " การตัดคำจากประโยคในภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด " Paper On Natural Language Processing, 2530-2537.

13. สมนึก คีรีโต " ระบบฐานความรอบรู้เพื่อการวิเคราะห์การแบ่งแยกพยางค์ไทย " รายงานการประชุมทางวิชาการวิศวกรรมไฟฟ้า สถาบันอุดมศึกษาแห่งประเทศไทย ครั้งที่ 9 , ธันวาคม 2529.

14. อนันต์ อ่วมศาสตร์และ เนาวรัตน์ อ่วมศาสตร์ ลักษณะภาษาไทย, สำนักพิมพ์ไทยวัฒนาพานิช , 2519.

15. Collins, William J. Data Structures An Object-Oriented Approach. Addison-Wesley Publishing 1992.

16. Gonnet, G.H. and R. Baeza-Yates. Handbook Of Algorithms And Data Structures in Pascal and C. 2nd.ed, Addison-Weseley, 1991.

17. Hoch, Rainer and Thomas Kiceinger. " On Virtual Partitioning of Large Dictionaries for Contextual Post - Processing to Improve Character Recognition. " IEEE. Conference 1993, pp. 226-231.

18. Kernighan, Brian W. and Dennis M. Ritchie. The C Programming Language, 2nd.ed. Prentice Hall, 1988.

19. Lewis, Harry R. and Larry Denenberg. Data Structures & Their Algorithms. HarperCollins Publisher, 1991.

20. Peterson, James L. " Computer Programs for Detecting and Corrceting Spelling Errors." Communication of the ACM , No. 12 (December 1980) , Vol. 23, pp. 676-687.

21. Peterson, James L. " A Note On Undetected Typing Errors. " Communication of the ACM, No. 7 (July 1986), Vol. 29, pp. 633-637.

22. S. Charnyapornpong. A Thai Syallable Separation Algorithm . AIT Thesis, Faculty of Engineer, No. CA-83-10, 1983.

23. Schurmann, Jurgen. " A Multifont Word Recognition System for Postal Address Reading. " IEEE Transaction on Computers, No. 8 (August 1978), Vol. c - 27, pp. 721-732 .

24. Tenenbaum, Aaron M. and Moshe J. Angenstein. Data Structures Using Pascal. Prentice-Hall Inc. 1986.

25. Tremblay, Jean-Paul. and Paul G.Sorenson. An Introduction to Data Structures with Applications. 2nd.ed, McGRAW-HILL , 1984.

26. Turner, Lawrence E. Jr. Data Structures From Recipes To C. Wm.C.Brown, 1994.

27. Weiss, Mark Allen. Data Structures And Algorithm Analysis. Benjamin Cummings Publishing, 1992.

ภาคผนวก

ภาคผนวก ก

ตารางค่ารหัส ASCII

View by: Decimal

Name: nbspace
Key:a0160 Dec: 160

Unicode: 00A0
Hex: A0

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127
p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
		...	ๆ	๚	๛	+	๏	ๆ	๚	๛	+	๏			
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
	•	๚	๛	๛	๛	๛	๛	ๆ	๚	๏	+	๏			ฯ
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
	ก	ข	ฃ	ค	ฅ	ฆ	ง	จ	ฉ	ช	ซ	ฌ	ญ	ฎ	ฏ
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
ฐ	ฑ	ฒ	ณ	ด	ต	ถ	ท	ธ	น	บ	ป	ผ	ฝ	พ	ฟ
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
ภ	ม	ย	ร	ฤ	ล	ฦ	ว	ศ	ษ	ส	ห	ฬ	อ	ฮ	ฯ
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
ะ	ั	า	ำ	ิ	ี	ึ	ื	ุ	ู	ฺ					฿
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
เ	แ	โ	ใ	ไ	ๅ	ๆ	็	่	้	๊	๋	์	๎	๏	๏
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255
๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	——	—			®	

ภาคผนวก ข

ตารางรายการคำที่ MICROSOFT WORD 6.0 ไม่สามารถตัดแบ่งคำได้

กฎบัตรสหประ-	กระโถนฤๅษี	กล้องจุลทรรศน์	กากขยาก
ชาชาติ	กระทั้น	กล้องตุด	กากณึก
กฎยุทธวินัย	กระแทกกระทั้น	กล้องโทรทรรศน์	กาญจน
กบฏุด	กระแท้ม	กลับตาลปัตร	กามราค
กปณก	กระปุกหลุก	กลุ้มอกกลุ้มใจ	กายพันธน์
กรณีย	กระเปาะเหลาะ	กษัตราธิราช	การิตการก
กรรมกรณ์	กระเปี้ย	กษัตริยชาติ	การิตวาจก
กรรมการิณี	กระแป้	กษีรามพุ	กาลโยค
กรรมชรูป	กระผลาม	กะดวน	กาพปักษี
กรรมชวาต	กระเฟียด	กะมง	กาพวาต
กรรมวาจาจารย์	กระแมม	กะส้าหอย	กำเริบเสิบสาน
กรรมสัมปาทิก	กระยาคชวาง	กัณฐ	กินหู้
กรรมาธิการ	กระยาสังแวง	กัณฐชะ	กี
กระจิ๋ง	กระลาพิน	กัณฐัศ	กีฏ
กระจิ๋ม	กระเษมสานต์	กัณฐัศว์	กีฏวิทยา
กระโจมอก	กระเหม่า	กัณห	กึกกือ
กระเฉง	กระอ่วน	กัณหธรรม	กุกกุฏ
กระโชกโฮกฮาก	กระอักกระอ่วน	กัณหปักษ์	กุกกุร
กระซ้อ	กรัณฑ	กัปปิย	กุฏาคาร
กระซิกกระซี้	กริตย	กัปปิยการก	กุมภ
กระซี้	กริยาวิเศษณานุ-	กัปปิยภัณฑ์	กุมภการ
กระเซ้ากระซี้	ประโยค	กัปปิยโวหาร	กุมภัณฑ
กระดิ้ง	กรุณาทฤคุณ	กัมป	กุมภัณฑยักษ์
กระเดียม	กฤด	กัมปน	กุมารลพิตา
กระแดะกระแด๋	กฤตติกา	กัมปนาการ	กุลาตีอก
กระติ้ง	กฤตย	กัมปนาท	กุศโลบาย
กระเตี้ยม	กฤษฎาญชเลศ	กัลยาณ	กูฏาคาร

เก๋งพั้ง
เกรงอกเกรงใจ
เกลือเบสิก
เกลือยูเรต
เกศี
เกษตรและสห-
กรณ์
แกงบวด
แก๊ส
แก๊สไข่เน่า
แก๊สเฉื่อย
แก๊สน้ำตา
โกกิล
โกญจ
โกญจนาท
โกร๋นเกร๋น
ขมึงทึง
ขรรค
ขล้ง
ขลึง
ขลุกขลัก
ขลุกขลิก
ขวายขวน
ขัณฑ
ขัณฑสกร
ขัณฑสีมา
ขัตติย
ขัตติยมานะ
ข้าวโกบ
ข้าวบิณฑ์
ข้าวปุ้น
ข้าวเภา
ข้าวฮาง
ขิด
ขี้กะโล้โท้
ขี้เข็บ
ขี้ทูด
ขี้ปะติ๋ว
ขี้แย
ขี้เหร่
ขี้นพลับพลา
ขุนพัฒน์
ขุย
ขุยไผ่
ขุยอินทรีย์
ขู
เขม
เขมรไทรโยค
เขือม
แขยงแขงขน
ไข้ทรพิษ
คช
คชกรรม
คชนาม
คชลักษณ์
คชศาสตร์
คชสาร
คชสีห์
คณ
คณบดี
คต
คมน
ครรไล
ครอบครองปร-
ปักษ์
คร่ำเครอะ
คริสตัง
คริสเตียน
ครื้นครึก
ครุฑพ่าห์
คลอแคล
คล้อแคล้
คล่องแคล่ว
คลอโรฟิลล์
คลับคล้ายคลับ
คลา
คลายเคล่ง
ควีนสิริกิติ์
คห
คหกรรมศาสตร์
คหบดี
คหปตานี
คัคน
คันถ
คันถธุระ
คันธมาทน์
คันบวย
คับอกคับใจ
คัพภ
คัพภสาลี
คัมภีร
คัมภีรภาพ
ค่ายเยาวชน
คาราคาก่า
คำนูณ
คิง
คิลาน
คิลานปัจจัย
คิลานเภสัช
คุมนุม
คุยห
คุยหฐาน
คุยหประเทศ
คุยหรหัสย์
คู่ปรปักษ์
เคมีอินทรีย์
เคร่งครัด
เครดิตฟองซิเอร์
เคห
เคหศาสตร์
เคหเศรษฐศาสตร์
เคหสถาน
เคอย
แค็ตตาล็อก
แคลเซียมคาร์ไบด์
แคลเซียมไซคลา-
เมต
ฆ้องหุ่ย
งกเงิ่น
งงงวย

ง่วงงุน
งอมแงม
งาลั่ว
งูสวัด
เงินดาวน์
เงินสเตอร์ลิง
โงนเงน
จตุ
จตุกาลธาตุ
จตุตถ
จตุทิพยคันธา
จตุบท
จตุบริษัท
จตุปัจจัย
จตุปาริสุทธิศีล
จตุร
จตุรคูณ
จตุรงคนายก
จตุรงคประดับ
จตุรงคยมก
จตุรถ
จตุรพักตร์
จตุรพิธ
จตุรพิธพร
จตุรภุช
จตุรภูมิ
จตุรยุค
จตุลังคบาท
จตุโลกบาล
จรรมขัณฑ์
จริม
จริย
จริยธรรม
จริยวัตร
จริยศาสตร์
จริยศึกษา
จองคช
จะกรัจจะกราจ
จ๊ะเอ๋
จักรพรรดิ
จักรวรรดิ
จักรวรรดินิยม
จักรวาต
จัณฑ
จัณฑวาตา
จัตวาทัณฑี
จัตุร
จัตุรพักตร์
จัตุรภุช
จัตุรมุข
จันทนา
จันทร
จันทรกลา
จันทรกานต์
จันทรคติ
จันทรคราส
จันทรพิมพ์
จันทรมณฑล
จันทรวงศ์
จันทรเศขร
จันทรุปราคา
จับยี่กี
จัว
จาค
จาตุทสี
จาตุร
จาตุรงค
จาตุรงคสันนิบาต
จาตุรราชการ
จำนวนจินตภาพ
จำนวนตรรกยะ
จำนวนอตรรกยะ
จิตต
จิตตภาวนา
จิตตวิสุทธิ
จิตตานุปัสสนา
จิตรปทา
จินต
จินตกวี
จินตนาการ
จินตภาพ
จิร
จิรกาล
จิรัฐิติกาล
จุฑาธิปไตย
จุดยุทธศาสตร์
จุทสมสุรทิน
จุล
จุลกฐิน
จุลชีพ
จุลชีวัน
จุลชีวิน
จุลทรรศน์
จุลวรรค
จุลศักราช
จุลสาร
จุลินทรีย์
เจดีย
เจดียถาน
เจดียสถาน
เจตนารมณ์
เจติย
เจริญสมณธรรม
เจ้าขุนมุลนาย
เจ้าเซ็นเต้นต๋ำบุด
เจ้าประคู้น
เจ้าสัว
เจีย
เจี๊อย
เจี๊อยแจ้ว
แจ่ง
แจ้น
แจบ
แจ่ม
แจ่มแจ้ง
แจ่มใส
แจ๋วแหวว
ฉกษัตริย์
ฉกามาพจร
ฉกามาวจร

ฉง
ฉทวาร
ฉทานศาลา
ฉพีสติม
ฉศก
ฉ ันท
ฉ ันทลักษณ์
ฉ ันทวิลาส
ฉ ันทศาสตร์
ฉินท
ฉินทฤกษ์
ฉุ กเฉิน
ฉุ กะหุก
เฉไฉ
เฉยเมย
เฉลิมพระ
ชนมพรรษา
โฉ่งฉ่าง
ชค
ชงฆ
ชนม
ชนมพรรษา
ชนินทร์
ชเนตตี
ชมชัว
ชมพูนุท
ชลธารก
ชลธิศ
ชลธี
ชว
ชอื้อ
ชอุ่ม
ชัค
ชังฆ
ชังฆวิหาร
ช่างฟิต
ชาต
ชาตรูป
ชาตสระ
ชาวเล
ชีรณัคคิ
ชีว
ชีวเคมี
ชีวประวัติ
ชีวโลก
ชีววิทยา
ชุณห
ชุณหปักษ์
ชุษณ
ชุษณปักษ์
เชษฐ
เชษฐะ
โชติก
โชโหว
ซอกแซก
ซ่องแซ่ง
ซับซาบ
ซึมซาบ
เซลล์ทุติยภูมิ
โซเดียมซัลเฟต
โซเดียมไซคลา-
เมต
โซเดียมไฮโดร-
เจนกลูทาเมต
โซเดียมไฮโดร-
เจนซัลเฟต
โซรมศัสตราวุธ
ฌาปน
ฌาปนกิจ
ฌาปนสถาน
ญาณทัสนะ
ฐานกรณ์
ฐาปน
ฑังส
ฑาก
ดนุช
ดอกดั้ว
ดะหมัง
ดังรือ
ดามพ
ดามพวรรณ
ดีเปรสชัน
ดีอกดีใจ
ดุ่ม
ดุริย
ดุริยางค
ดุริยางคศาสตร์
ดุริยางคศิลป์
ดุลย
ดุลยพินิจ
ดุลยภาพ
ดูดู๋
ดูรึ
เด็กเมื่อวานซืน
เดโช
เดโชชัย
เดโชพล
แดะแด๋
ตกขุย
ตกใต้เถรเทวทัต
ตจ
ตจปัญจก
กรรมฐาน
ตจสาร
ตติย
ตนุมัธยา
ต้มโคล้ง
ต้มเปอะ
ตโมนุท
ตรมตรอม
ตรรก
ตรรกศาสตร์
ตรัง
ตรังค
ตรังควชิราวุดี
ตราไปรษณียากร
ตรีกันสวาต
ตรีฉินทลามกา
ตรีชาต
ตรีทิพยรส

ตรีทุรวสา
ตรีเทวตรีคันธา
ตรีปิตผล
ตรีภูว
ตรีมธุร
ตรีโลจน์
ตรีโลหก
ตรีวาตผล
ตรีสุรผล
ตรีเสมหผล
ตรึงศ
ตฤตีย
ตฤๅ
ต่อนัดต่อแนง
ตอม่อ
ตะเกียงแก๊ส
ตะพึดตะพือ
ตะหลุง
ตั้งอกตั้งใจ
ตัตวศาสตร์
ตันอกตันใจ
ตาโก้ง
ตามพ
ตามพหัตถี
ตายขุย
ตาลปัตร
ตาลปัตรบังเพลิง
ตาลปัตรฤาษี
ตาลปัตรฤาษี
ตาลอ

ติ
ติก
ติงส
ติงสติม
ติดตื๋อ
ตุ๊ต๊ะ
ตุ้บตั้บ
ตุรงค
ตุล
ตุหรัดตุเหร่
เต็มอก
เต็มอกเต็มใจ
เตาแก๊ส
เต้าหู้ยี้
แตรฟันฟาร์
แตรวงโยธวาทิต
ไตรกิศยา
ไตรตรึงษ์
ไตรยางศ์
ไตรวัฏ
ไตรสรณาคมน์
ถนัดถนี่
ถมถืด
ถมเถ
ถวัดถวัน
ถั่วคร้า
ถ่านโค้ก
ถ่านไฟแช็ก
ถิร
ถีน

ถึงแก่มรณภาพ
ถึงแก่อสัญกรรม
ถือน้ำพระ
พิพัฒน์-สัตยา
ถุล
ถูป
ถูปารหบุคคล
ถูล
เถย
เถยจิต
เถยเจตนา
เถยสังวาส
เถลไถล
ไถย
ทนต
ทนตกาษฐ์
ทร
ทรกรรม
ทรงข้าวบิณฑ์
ทรงมัณฑ์
ทรชน
ทรชาติ
ทรทึง
ทรธึก
ทรพล
ทรพิษ
ทรพี
ทรภิกษ์
ทรยศ
ทรยุค

ทรรศนาการ
ทรราช
ทรลักษณ์
ทรัพย
ทรัพยสิทธิ
ทฤฆ
ทลบม
ทวัตดึงสาการ
ทวาทศม
ทวาทศี
ทวาบร
ทวิช
ทวิชงค์
ทวิป
ทศเบญจกูล
ทศพิธราชธรรม
ทศม
ทองคำเปลว
ทองเค
ทองชมพูนุท
ทองบรอนซ์
ทองสักโก
ท้องอัสดงคต
ทอดหุ่ย
ทักษ
ทัณฑ
ทัณฑกรรม
ทัณฑฆาต
ทัณฑนิคม
ทัณฑวิทยา

ทัณฑสถาน
ทัต
ทันต
ทันตชะ
ทันตแพทย์
ทัศน
ทัศนคติ
ทัศนวิสัย
ทัศนศาสตร์
ทัศนศึกษา
ทัศนีย
ทัศนียภาพ
ทาร
ทิชากร
ทิฐ
ทิพย
ทิพยจักษุ
ทิพยจักษุญาณ
ทิพยญาณ
ทิพยเนตร
ทิพยพยาน
ทิพยมานุษ
ทิพยรส
ทิพยโสตร
ทิวส
ทีฆ
ทีฆชาติ
ทีฆนิกาย
ทีฆสระ
ทีฆายุ
ทีรฆ
ทุกข
ทุกขนิโรธ
ทุกขลาภ
ทุกขเวทนา
ทุกขสมุทัย
ทุกขารมณ์
ทุติย
ทุร
ทุรกันดาร
ทุรคม
ทุรชน
ทุรชาติ
ทุรพล
ทุรภิกษ์
ทุรลักษณ์
ทุษฐ
ทุส
ทูร
เทพย
เทพยเจ้า
เทพยดา
เทเพนทร์
เทลลูเรียม
เทว
เทวทัณฑ์
เทวทูต
เทวธรรม
เทวธิดา
เทวนาครี
เทวนิยม
เทวภาวะ
เทวรูป
เทวโลก
เทววาจิกะ
เทววิทยา
เทวสถาน
เทห์
เทห์ฟากฟ้า
เทียรฆ
เทียรฆชาติ
เทียรฆราตร
แท็กซี่มิเตอร์
โทรทรรศน์
โทรมัศตราวุธ
โทษกรณ์
โทษารมณ์
โทส
ธงพระครุฑพ่าห์
ธงเยาวราชน้อย
ธงเยาวราชใหญ่
ธนาณัติ
ธนิษฐะ
ธนุร
ธนุรมารค
ธนุรเวท
ธรณีศวร
ธรรมกถึก
ธรรมนูญ
ธรรมนูญการปกครองราชอาณาจักร
ธรรมะธัมโม
ธรรมาภิมุข
ธราธิบดี
ธราธิป
ธเรษตรีศวร
ธัญดัจ
ธาตุโขภ
ธานินทร์
ธาราธิคุณ
ธีร
ธีรภาพ
ธีรราช
ธุดงค
ธุดงควัตร
ธุดงคสมาทาน
ธุม
ธุมเกตุ
ธุมชาล
ธุมเพลิง
ธุมา
ธุร
ธุรการ
ธุว
ธุวดารา
ธุวภาค
ธุวมณฑล
ธุวยัษฎี

ธูปฤาษี
นข
นขทารณ์
นขลิขิต
นขเลขา
นขา
นค
นภ
นภจร
นภดล
นภมณฑล
นภวิถี
นภศูล
นภสินธุ์
นมบกอกพร่อง
นย
นร
นรเทพ
นรนาถ
นรนายก
นรบดี
นรบาล
นรพยัคฆ์
นรราช
นรเศรษฐ์
นรสิงห์
นรสีห์
นฤ
นฤคหิต
นฤโมษ
นฤดม
นฤตย
นฤตย์
นฤตยศาลา
นฤตยศาสตร์
นฤตยสถาน
นฤเทพ
นฤนาท
นฤบดี
นฤบาล
นฤเบศ
นฤป
นฤปนีติ
นฤปเวศม์
นฤปัตนี
นฤพาน
นฤมล
นฤมาณ
นฤมิต
นว
นวก
นวกภูมิ
นวกรรม
นวกรรมิก
นวการ
นวกิจ
นวครหะ
นวคุณ
นวเคราะห์
นวชาต
นวทวาร
นวนิต
นวนิยาย
นวปฏล
นวพธู
นวมุข
นวรัตน์
นวโลหะ
นววิธ
นวอรหาทิคุณ
นวารหาทิคุณ
นักเลง
นักษัตรเนมี
นักษัตรปาฐก
นักษัตรโยค
นักสราช
นันท
นันทปักษี
นัวเนีย
นาคสังวัจฉร
นาคินทร์
นาฏย
นามไธย
นามาภิไธย
น้ำพระพิพัฒน์-
สัตยา
นิคมสหกรณ์
นิคมอุตสาหกรรม
นิคห
นิคหกรรม
นิตย
นิตยทาน
นิตยภัต
นิตยสาร
นิติกรณ์
นิทรารมณ์
นิทาฆ
นิพัทธ
นิพัทธกุศล
นิยยานิก
นิรโฆษ
นิรย
นิรยบาล
นิรันตร
นิศาทิ
นิสภ
นีร
นีรจร
นีรช
นุ่งกระโจมอก
นุงนัง
เนาว
เนาวรัตน์
แน่นนันต์
แน่นแฟ้น
แน่นหน้าอก
บงก
บงสุกูลิก
บดีวรดา
บทภาชน์

บทภาชนีย์
บทเรศ
บทวเรศ
บมิ
บร
บรทาร
บรทารกรรม
บรรพตกีลา
บรอนซ์
บราง
บริษัทเครดิต-
ฟองซิเอร์
บแรง
บแรงภักษ์
บอกหนังสือ-
สังฆราช
บอง
บเอ
บัญญัติไตรยางศ์
บัฐยาพฤต
บัณเฑาะก์
บัดแมล่ง
บาทนิเกต
บาทภัฏ
บาทรช
บานเบียง
บานโภชน์
บานีโยทก
บาร์เรล
บ้าระห่ำ
บำเรอเชอภักดิ์
บิณฑ
บิณฑบาต
บิตุจฉา
บุบบิบ
บุปผ
บุปผชาติ
บุปผวิกัติ
บุพเปตพลี
บุพพ
บุพพาษาฒ
บุพพาสาฬหะ
บุพวิเทห์
บุรพผลคุนี
บุราณทุติยิกา
บุษบ
บุษบราค
บุษบวรรษ
บุษบัน
บุษบาบัณ
บุษป
บุษยะ
บุหงาประหงัน
บุหงามลาซอ
บูชนีย
บูชนียสถาน
บูรณ
บูรณภาพ
บูรณมี
บูรพาษาฒ
เบญจกูล
เบญจคัพย์
เบญจม
เบญจมสุรทิน
เบญจโลหกะ
เบญจวรรณ
เบญจวรรณห้าสี
เบญจางค
เบญจางค
ประดิษฐ์
เบอะบะ
แบะอก
โบกขร
โบกขรพรรษ
ใบพัทธสีมา
ปฏิ
ปฏิการ
ปฏิคหิต
ปฏิญญาสากลว่า-
ด้วยสิทธิมนุษย-
ชน
ปฐมโพธิกาล
ปฐมสุรทิน
ปฐมาษาฒ
ปนเป
ปปัญจ
ปปัญจธรรม
ปโย
ปโยชนม์
ปโยธร
ปโยธรา
ปโยนิธิ
ปโยราศิ
ปร
ปรตยักษ์
ปรนัย
ปรปักษ์
ปรม
ปรโลก
ปรวาที
ประกายพรึก
ประคบประหงม
ประโคนธรรพ์
ประจักษ
ประจักษนิยม
ประเจิดประเจ้อ
ประชาธิปไตย
ประชาภิบาล
ประดักประเดิด
ประดิ
ประดิษฐ
ประดิษฐกรรม
ประเดี๋ยวประด๋าว
ประทุษฏจิต
ประทุษฐจิต
ประมุท
ประลัยวาต
ประสัยห
ปรางค์ปรา
ปริณามัคคิ

ปริย
ปริยรณ
ปริยวาท
ปริยวาที
ปริเวณ
ปรู๊ดปร๊าด
ปลวังค
ปลาม้ำ
ปวัตน
ปวัตนาการ
ปสันน
ปสันนาการ
ปสัยห
ปสัยหาการ
ปสัยหาวหาร
ปะวะหลิ่ม
ปักข
ปักขคณนา
ปักขพิพาร
ปักษ
ปักษกษัย
ปักษคณนา
ปักษคม
ปักษธร
ปักษเภท
ปักษวาหน
ปักษานดร
ปักษาวสาน
ปักษ
ปัจจันต
ปัจจันตคาม
ปัจจันตชนบท
ปัจจันตประเทศ
ปัจจุส
ปัจจุสกาล
ปัจจุสมัย
ปัจเจกโพธิ
ปัจโจสกาล
ปัจยาการ
ปัญจ
ปัญจนที
ปัญจม
ปัญจวัคคีย์
ปัญจสาขา
ปัญญาส
ปัณรส
ปัณรสม
ปัณรสมสุรทิน
ปัทม
ปัทมปาณี
ปัทมราค
ปัทมราช
ปันจุเหร็จ
ปาฏิหาริย์
ปาฐ
ปาฐก
ปาฐกถา
ปาณ
ปาณก
ปาณกชาติ
ปาณทัณฑ์
ปาณนาศ
ปาณภูต
ปาณวินาศ
ปาณสาร
ปาณิมุกต์
ปานีย
ปาป
ปาปมุต
ปาร
ปารคู
ป่าเส็งเคริ่ง
ปาหุไณย
ปาหุไณยบุคคล
ปิงคล
ปิงคลหัตถี
ปีตามหะ
ปีตามหัยกา
ปีตามหัยยิกา
ปีโตรเลียม
ปีย
ปีหลั่น
ปีมะโว้
ปุนัพพสู
ปุพพ
ปุพพผลคุนี
ปุพพะภัททะ
ปุม
ปุ่มป่ำ
ปุรพผลคุนี
ปุรพษาฒ
ปุษยะ
ปุสสะ
ปูชนีย
ปู่ยี่ปู่ย่ำ
เปกข์
เปิดก่า
เปต
เปตพลี
เปตวิสัย
เปตอง
เปรมปรา
เปลว
เปศลมัธย์
เปิดอก
เปี๊ยะ
แปรปรวน
แป๊ะ
โปกขร
โปกขรพรรษ
โป้งโย้ง
โปฐบท
โปฐปทมาส
โปร่งเปร่ง
ไปรษณีย
ไปรษณียนิเทศ
ไปรษณียบรรณ
ไปรษณียบัตร
ไปรษณียภัณฑ์
ไปรษณียวัตถุ

ไปรษณียากร
ผมชิงเกิล
ผมบ๊อบ
ผรณ
ผรุส
ผรุสวาท
ผลคุน
ผลคุนี
ผลคุนีบูรพมาส
ผลับ
ผลิก
ผลิกศิขรี
ผลิกศิลา
ผลุนผลัน
ผลุบผลับ
ผสมผเส
ผัสส
ผากแผ้ง
ผาลคุน
ผิน
ผินหลังให้
ผีต
ผูกพัทธสีมา
เผน
เผนธรรม
เผยิบผยาบ
เผลอไผล
เผือกโทป้าด
โผฏฐัพ
โผฏฐัพธรรม
โผฏฐัพพะ
โผย
โผอน
ฝรั่งเศส
พงศ
พงศกร
พงศธร
พจน
พญาพยาต
พณิช
พณิชย์
พนขัณฑ์
พยักพเยิด
พยัคฆ
พยุห
พยุหบาตร
พยุหบาตรา
พยุหยาตรา
พยุหโยธา
พยุหเสนา
พยุหแสนยา
พยุหแสนยากร
พรรคกลิน
พรหมจรรย์
พรหมจาริณี
พร้อมพรัก
พระทัย
พระฤาษี
พระหารณย์
พระห่ารัณย์
พฤกษ
พฤกษชาติ
พฤกษทล
พฤกษเทวดา
พฤกษราช
พฤกษศาสตร์
พฤฒ
พฤฒา
พฤฒาจารย์
พลว
พลังงานจลน์
พลับพลา
พอโลเนียม
พักตร
พัฒน
พัทธ
พัทธสีมา
พันธ
พันธกรณี
พันธบัตร
พันธมิตร
พันวรรษา
พ่าห์
พิจิก
พิถย
พิถี
พิทย
พิทยาคาร
พิพัฒน์
พิพิธภัณฑสถาน
พิมพ
พิมพาภรณ์
พิริย
พิริยพฤนท์
พิริยโยธา
พิลึกกึกกือ
พิลึกพิลั่น
พี่เบิ้ม
พีร
พุทธมามกะ
พุทธศาสนิกชน
พุ่มข้าวบิณฑ์
เพ้อเจ้อ
แพทย
แพทยศาสตร์
โพธิ
โพธิญาณ
โพธิบัลลังก์
โพธิปักขิยธรรม
โพธิสมภาร
โพธิสัตว์
โพนเพน
โพยก๊วน
ไพ่ป๊อก
ไพรสาณฑ์
ฟอดแฟด
ฟัดเฟียด
เฟอะฟะ
เฟะฟะ
ไฟแช็ก

ภย
ภราตร
ภราตรภาพ
ภราตฺฤ
ภราตฺฤภาพ
ภวังค
ภวังคจิต
ภักต
ภักตกฤตย์
ภักษ
ภักษฺการ
ภังค
ภัณฑ
ภัณฑครรภ
ภาคเสธ
ภาชนีย
ภาระติดพันใน-
อสังหาริมทรัพย์
ภาว
ภาวศุทธิ
ภาวะฉุกเฉิน
ภิกษุ
ภิน
ภินชาติ
ภินท
ภินทน
ภินวรรณ
ภีรุก
ภีรุกชาติ
ภุกต
ภุกตเศษ
ภุญช
ภูธเรศวร
ภูเนตฺ
ภูเบนทร์
ภูเบศ
ภูเบศวร์
ภูป
ภูมินทร์
ภูมิบริมาณ
ภูมิศาสตร์
เศรษฐกิจ
ภูมีศวร
ภูรโลก
ภูว
ภูวดล
ภูวนัตตรัย
ภูวนาถ
ภูวเนตร
ภูวไนย
ภูษณ
ภูษณพาส
เภสัชพฤกษ-
ศาสตร์
เภสัชอุตสาห-
กรรม
โภชน
มณีราค
มต
มธุป
มธุร
มธุรตรัย
มธุรพจน์
มธุลีห์
มนท
มนทกานติ
มนัสดาป
มนุษย
มนุษยชาติ
มนุษยเทพ
มนุษยธรรม
มนุษยโลก
มนุษยศาสตร์
มนุษยสัมพันธ์
มนุษย์อวกาศ
มโนนุกูล
มโนภินิเวศ
มยุร
มยุรคติ
มยุรฉัตร
มยุรอาสน์
มรณ
มรณกรรม
มรณธรรม
มรณบัตร
มรณภัย
มรณภาพ
มรณันติก
มฤคศิร
มฤคศิรมาส
มฤคศิรัส
ม้วนต้วน
มหรณพ
มหัจฉริย
มหันต
มหันตโทษ
มหิป
มอมแมม
ม่อห้อม
มะม่วงหิมพานต์
มะสัง
มัค
มัคนายก
มังคละ
มังส
มังสวิรัติ
มัช
มัชฌ
มัชฌันติก
มัชฌันติกสมัย
มัชฌิม
มัชฌิมนิกาย
มัชฌิมบุรุษ
มัชฌิมประเทศ
มัชฌิมภูมิ
มัชฌิมยาม
มัชฌิมวัย
มัชวิรัติ
มัณฑ
มัณฑน

มัณฑนศิลป์
มัธย
มัธยฐาน
มัวซัว
มาคสิร
มาฆ
มาฆบูชา
มานทะลุน
มานุษย
มานุษยวิทยา
มิคสิร
มิจฉากัมมันตะ
มิญช
มิต
มิตภาณี
มิ่ม
มิศรก
มิส
มีพห
มุขบาฐ
มุขย
มุขยประโยค
มุตกิด
มุทธาภิเษก
มุบมิบ
มุรธาภิเษก
มุฮัมมัด
มูค
มูรธ
มูละ
มูสัง
มูพห
เมรย
แม่น
แม่นยำ
แมลงวันสเปน
โมกข
โมกขบริสุทธิ์
โมกษ
โมฆ
โมฆกรรม
โมฆียะ
โมฆียกรรม
โมร
โมห
ไม่กี่อัฐพส
ไม้ทัณฑฆาต
ยติภังค์
ยมกปาฏิหาริย์
ยว
ยวา
ยั้ง
ยัดเยียด
ยันตร
ยันตรกรรม
ยางมะตอย
ยาปนมัต
ยามโยค
ยิมนาสติก
ยี้
ยุคันต
ยุคันตวาต
ยุทธ
ยุทธการ
ยุทธนาธิการ
ยุทธภัณฑ์
ยุทธภูมิ
ยุทธโยธา
ยุทธวิธี
ยุทธวินัย
ยุทธศาสตร์
ยุทธหัตถี
ยุทโธปกรณ์
ยุพ
ยุพดี
ยุพยง
ยุพเยาว์
ยุว
ยุวชน
ยุวราช
ยุวราชา
เย็บอก
เยภุย
เยภุยนัย
เยภุยสิกา
เยาว
เยาวชน
เยาวนะ
เยาวมาลย์
เยาวยอด
เยาวราช
เยาวเรศ
เยาวลักษณ์
เยิ่นเย้อ
โยกเยก
โยค
โยคจักษุ
โยคนิทรา
โย่งเย่ง
โยธ
โยธวาทิต
โยพนมัท
รกเรี้ยว
รงค
รงควัตถุ
รถฉุกเฉิน
รมณีย
รวนเร
รโหคต
รังสรัง
รังสีคอสมิก
รัชด
รัชต
รัฐธรรมนูญ
รัฐประศาสนนัย
รัฐประศาสน-
ศาสตร์
รัฐประศาสโน-
บาย
รัตน

รัตนโกสินทร์
รัตนตรัย
รัตนบัลลังก์
รัตนสิงหาสน์
รัมก
รัมณีย
รัส
รัสสระ
ราชกิจจานุเบกษา
ราชภัฏ
ราชมัล
ราชโยค
ราชวโรงการ
ราชหัตถเลขา
ราชองครักษ์
ร้ายกาจ
รุกข
รุกขชาติ
รุกขมูล
รุกขมูลิกธุดงค์
เรียบวุธ
เรื่อยเจื้อย
แรงงานและ-
สวัสดิการสังคม
ฤๅทัย
ฤๅษี
ฤๅษีแปลงสาร
ฤๅษีผสมแล้ว
ฤๅษีเลี้ยงลิง
ลฆุโภชน์
ลหุก
ลองธรรม์
ลักษณ
ลักษณนาม
ลักษณาการ
ลักษมณ์
ลัคน
ลัยคต
ลายสือ
ลิต
ลุ่นตุ้น
ลุ่นโตง
ลูกมโหตร
เลขาธิการ
เล่นพิเรนทร์
เลป
เลปกร
เลอะเทอะ
เละเทะ
เลิ่กลั่ก
เลือดในอก
เลื้อยเจื้อย
โลกเชษฐ์
โลกาภิวัตน์
โลกุตระ
โลห
โลหกุมภี
ไล่เลี่ย
ฦๅ
ฦๅชา
ฦๅสาย
วงศ
วงศกร
วชิร
วชิรปาณี
วชิรหัตถ์
วณ
วณบัตร
วณพันธน์
วต
วทานิย
วธ
วธก
วร
วรดนู
วรท
วรทาน
วรมหาวิหาร
วรรณ
วรรณกรรม
วรรณคดี
วรรณยุกต์
วรรณยุต
วรรณศิลป์
วรรษา
วรวิหาร
วสันต
วสันตวิษุวัต
วัคคุวัท
วัจ
วัจกุฎี
วัจมรรค
วัชร
วัชรธาตุมณฑล
วัชรปาณี
วัชรยาน
วัชรอาสน์
วัฏ
วัฏจักร
วัฏทุกข์
วัฏสงสาร
วัฒน
วัฒนธรรม
วันทย
วันทยหัตถ์
วัส
วัสดณนา
ว้าก
วากจิรพัสตร์
วากย
วากยสัมพันธ์
วาต
วาตปานะ
วาตภัย
ว่าแต่เขาอิเหนา-
เป็นเอง
วาทย
วาทยกร
วานซืน
วาลธิ

วิกขัมภนะ
วิกัติ
วิกัติการก
วิกาลโภชน์
วิ่งระแบง
วิทย
วิทยฐานะ
วิทยาคาร
วิทยาศาสตร์ชีว-
ภาพ
วินย
วิร
วิโรฒ
วิโรธน์
วิลย
วิไลวรรณ
วิวรรธน์
วิวัฒน
วิวาห
วิศาข
วิศาขบูชา
วิเศษณ
วิเศษณการก
วิสย
วิสาข
วิสาขบูชา
วิสามานยนาม
วีต
วีร
วีรกรรม

วีรชน
วีรบุรุษ
วุ่นเป็นจุลกฐิน
เวนคืนอสังหาริม-
ทรัพย์
เวมัติก
เวียง
ศตปที
ศตสังวัตสร์
ศระวณ
ศราทธ
ศราทธพรต
ศศ
ศศธร
ศศพินฑุ์
ศศลักษณ์
ศศิขัณฑ์
ศักย
ศักยภาพ
ศักร
ศักรภพน์
ศัพท
ศัพทมูลวิทยา
ศัลย
ศัลยกรรม
ศัลยแพทย์
ศัลยศาสตร์
ศัสตร
ศัสตรกรรม
ศัสตรการ

ศัสตรศาสตร์
ศากย
ศากยเกตุ
ศากยพุทธ
ศากยมุนี
ศานต
ศานตรส
ศานติก
ศาปมุกติ์
ศาลคดีเด็กและ-
เยาวชน
ศาลเยาวชนและ-
ครอบครัว
ศาสตร
ศาสน
ศาสนจักร
ศาสนีย
ศิร
ศิรประภา
ศิลป
ศิลปกร
ศิลปกรรม
ศิลปการ
ศิลปกิจ
ศิลปวิทยา
ศิลปศาสตร์
ศิว
ศิวโมกข์
ศิวลึงค์
ศิวเวท

ศิษฏ์
ศีต
ศีตกาล
ศึกษาธิการ
ศึกษานิเทศก์
ศุกร
ศุกรวรรณ
ศุกล
ศุกลปักษ์
ศุภ
ศุภกร
ศุภเคราะห์
ศุภนิมิต
ศุภมัสดุ
ศุภมาตรา
ศุภมาส
ศุภร
ศุภอักษร
ศุลก
ศูนย
ศูนย์พ่าห์
ศูนยภาพ
ศูนย์เยาวชน
ศูนยวาท
เศรษฐ
เศรษฐกิจ
เศรษฐศาสตร์
เศาจ
โศกศัลย์
ษมายุมแปลง

สกฎ
สกฎภาร
สกลมหาสังฆ
ปริณายก
สงสารวัฏ
สตัมภ์
สถวิร
สถวีร
สถิร
สนุกเกอร์
สภาว
สภาวการณ์
สมณ
สมณบริขาร
สมณโวหาร
สมณศักดิ์
สมณสารูป
สมณสาสน์
สมณะ
สมเด็จพระสังฆ-
ราช
สมถ
สมถยานิก
สมถวิปัสสนา
สมบัตินักเลง
สมปฤๅดี
สมฤๅดี
สมิติ
สมุห
สมุหกลาโหม
สมุหเทศาภิบาล
สมุหนาม
สมุหนายก
สมุหประธาน
สมุหพระกลาโหม
สรณ
สรณคมน์
สรณตรัย
สรณาคมน์
สรีร
สรีรกิจ
สรีรธาตุ
สรีรวิทยา
สรีรศาสตร์
สลบไสล
สลาง
สว
สวกรรม
สวการย์
สวนีย
สวภาพ
สวรรค
สวรรคต
สวรรคบดี
สวรรค์อยู่ในอก-
นรกอยู่ในใจ
สวราชย์
สวัสดิการ
สวัสติ
สวามินี
สวามิภักดิ์
สห
สหกรณ์
สหการ
สหจร
สหชาต
สหชาติ
สหธรรม
สหธรรมิก
สหบาน
สหประชาชาติ
สหพันธ์
สหพันธรัฐ
สหภาพ
สหภาพแรงงาน
สหศึกษา
สหัส
สหัสธารา
สหัสนัยน์
สหัสเนตร
สหัสรังสี
สหัสสะ
สอนหนังสือ-
สังฆราช
สะบักสะบอม
สะบัดสะบิ้ง
สังข
สังขกร
สังขต
สังขตธรรม
สังค
สังคญาติ
สังฆ
สังฆกรรม
สังฆการี
สังฆเถระ
สังฆทาน
สังฆนายก
สังฆปาโมกข์
สังฆภัต
สังฆเภท
สังฆมณฑล
สังฆมนตรี
สังฆราช
สังฆสภา
สังฆาธิการ
สัชฌ
สัชฌกร
สัชฌการ
สัญชาต
สัญชาตญาณ
สัตตาห
สัตถ
สัตม
สัตมวาร
สัตย
สัตยพรต
สัตยวาที
สัตว
สัตวบาล

สัตวแพทย์
สัตวโลก
สัตววิทยา
สัทธาธิก
สันถว
สันถวไมตรี
สันทิฐิก
สับปลี้
สัปด
สัปดปกรณ์
สัปดประกรณ์
สัปดสดก
สัปต
สัปตศก
สัปปุริส
สัมพัจฉร
สัมพัจฉรฉิ นท์
สัมพันธ
สัมพันธภาพ
สัมพันธมิตร
สัมพันธไมตรี
สัมพาหะ
สัมพุทธ
สัมมากัมมันตะ
สัมฤทธิ
สัมฤทธิศก
สัลลาป
สัลเลข
สัสต
สัสตทิฐิ
สากษิน
สากหยาก
สากิย
สากิยบุตร
สากิยมุนี
สาง
สาฎิก
สาณ
สาธย
สามนต
สามนตราช
สามานย
สามานยนาม
สารบบ
สารบับ
สารัตถ
สาราณีย
สาราณียธรรม
สาหัตถ
สาหัตถิก
สำนึง
สิขานล
สิงห
สิงหนาท
สิงหบัญชร
สิงหราช
สิงหลก
สิญจ
สิทธ
สิร
สีฆ
สีต
สีห
สีหนาท
สีหบัญชร
สีหราช
สุคนธ
สุคนธรส
สุญญ
สุทธ
สุธาโภชน์
สุนทรีย
สุพรรณภาชน์
สุภ
สุ้มเสียง
สุร
สุรคต
สุรเชษฐ์
สุรบถ
สุรภาพ
สุรโลก
สุรสีหนาท
สุรเสียง
สุริย
สุริยกันต์
สุริยกานต์
สุริยการ
สุริยกาล
สุริยคติ
สุริยคราส
สุริยเคราะห์
สุริยมณฑล
สูป
สูรย
สูรยกานต์
สูรยคราส
เสกข
เสกขบุคคล
เสข
เสขบุคคล
เสิ่งเคริ่ง
เสต
เสนห
เสนาธิการ
เสนาธิปัต
เสนาพยุห์
เสนาพยูห์
เสล
เสลบรรพต
เสารภย์
เสาว
เสาวคันธ์
เสาวธาร
เสาวภา
เสาวภาคย์
เสาวภาพ
เสาวรภย์
เสาวลักษณ์
เสือบอง
แสนย

โสตทัศนวัสดุ
โสตทัศนอุปกรณ์
โสภ
หงส
หงสคติ
หงสบาท
หงสรถ
หงสราช
หงำเหงอะ
หญ้ายายเภา
หทย
หน็องแหน็ง
หนอนตายหยาก
หน้าอก
หมักหมม
หมากฮอส
หม้ำ
หม้ำตับ
หมุบหมิบ
หมูหย็อง
หย
หยอกเอิน
หย็อง
หย็องกรอด
หย็องแหย็ง
หย็อมแหย็ม
หย็อย
หยิกหย็อง
หรณ
หรรษ
หรัสว
หรัสวมูรดี
หฤทย์
หฤทัยกัปน์
หฤทัยกัมป์
หฤทัยปรีย์
หฤทัยพันธน์
หลงใหล
หลาท
หลืบ
หลู่หลี่
หว่าง
หสน
หอมเตียม
หักอก
หักอกหักใจ
หังส
หัตถ
หัตถกรรม
หัตถการ
หัตถกิจ
หัตถบาส
หัตถพันธ์
หัตถศิลป์
หัตถศึกษา
หัตถาจารย์
หัตถาภรณ์
หัวหน่าว
หัวอก
หัสต
หิม
หิมพาน
หิมพานต์
หิมวัต
หิมวันต์
หิมวา
หิมวาต
หิมวาน
หิมวาส
หิมเวศ
หิรัณย
หิรัณยการ
หิรัณยเกศ
หิรัณยรัศมี
หีน
หีนยาน
หึงส
หื่นหรรษ์
หือรือโหด
หุต
หุตโกวิท
หู่หี่
เหมันต
เหมาะเหม็ง
แหย็ม
ไหหลำ
อก
อกไก่
อกคราก
อกตั้ง
อกเต่า
อกแตก
อกทะเล
อกนิษฐ์
อกเมือง
อกรวบ
อกโรย
อกเลา
อกหัก
อกัปปิย
อกัปปิยวัตถุ
อกัปปิยโวหาร
อคาร
องค
องค์การสห-
ประชาชาติ
องคชาต
องครักษ์
องควิการ
องควิเกษป
องควิทยา
อจล
อจิร
อดุลย
อทระ
อธิก
อธิกมาส
อธิกวาร
อธิกสุรทิน
อธิปไตย

อน
อนงค
อนงคเลข
อนงคเลขา
อนธ
อนธการ
อนัตถ
อนันต
อนันตร
อนากูล
อนารย
อนารยชน
อนารยธรรม
อนึก
อนึกทรรศนะ
อนึกทัศนะ
อนุตร
อนุรักษ
อนุรักษนิยม
อนุราช
อนุราชะ
อเนกอนันตัง
อปจายน
อปจายนธรรม
อปจายนมัย
อปร
อปรภาค
อปรัณณชาติ
อปาน
อภิชนาธิปไตย
อภิณห
อมต
อมตบท
อย
อยน
อรทัย
อรรถกถาจารย์
อรหันต
อรหันตฆาต
อรัญญิก
อรัญวาส
อริย
อริยทรัพย์
อริยบุคคล
อริยผล
อริยมรรค
อริยสัจ
อวิญญาณก
อวิญญาณกทรัพย์
อวิโรธน์
อสังหาริม
อสังหาริมทรัพย์
อสัญ
อสัญกรรม
อสัมภิน
อสัมภินพงศ์
อสัมภินวงศ์
อสิเลสะ
อสุร
อสุรกาย
อเสกข
อเสกขบุคคล
อเสข
อเสขะ
อ่อนอกอ่อนใจ
อ้อมอก
อ้อยส้อย
อะดรีนาลีน
อะดักอะเดี้ย
อะดักอะเดื่อ
อะดักอะแด้
อักขร
อักขรวิธี
อักขรวิบัติ
อัคนิพ่าห์
อัคนิโหตร
อัคร
อัครชายา
อัครมหาเสนาบดี
อัครมเหสี
อัครราชทูต
อัครสมณทูต
อังส
อังสกุฏ
อัจฉริย
อัจฉริยบุคคล
อัจฉริยภาพ
อัจฉริยลักษณ์
อัจฉริยลักษณะ
อัชฌัตติก
อัญ
อัญดิตถีย์
อัญเดียรถีย์
อัญมณี
อัญมัญ
อัฏฐ
อัฐม
อัฐพส
อัฐิมิญชะ
อัฒ
อัฒจันทร์
อัฒภาค
อัฒมณฑล
อัฒมาส
อัฒรัตติ
อัณฑ
อัณฑโกส
อัณฑชะ
อัต
อัตชีวประวัติ
อัตตาธิปไตย
อัตนัย
อัตภาพ
อัตวินิบาตกรรม
อัตหิต
อัตเหตุ
อัทธ
อัทธคต
อัทธคู
อันต

อันตกาล
อันตกิริยา
อันตคุณ
อันตคู
อันตชาติ
อันตลิกขะ
อั้นตั้น
อันติม
อันติมสัจ
อันเตวาสิก
อันธ
อันธการ
อันธพาล
อันโยนย
อันวย
อัป
อัปเปหิ
อัพภันตร
อัมพ
อัมพวัน
อัมพวา
อัมพุช
อัมพุชินี
อัมพุท
อัลกุรอาน
อัศว
อัศวโกวิท
อัศวมุข
อัศวมุขี
อัศวเมธ
อัศวยุช
อัษฎ
อัษฎกฉัน์
อัษฎมงคล
อัษฎมงคล
อัส
อัสดงคต
อัสดร
อัสมุขี
อัสสนี
อาคันตุก
อาคันตุกภัต
อาคันตุกวัตร
อาจริย
อาจริยวัตร
อาจริยวาท
อาชีว
อาชีวศึกษา
อาทิตย
อาทิตยมณฑล
อาทิตยวาร
อานาปานัสสติ
อาปณกะ
อ้ายเงี๊ยว
อายุร
อายุรแพทย์
อายุรเวช
อายุรเวท
อายุรศาสตร์
อารย
อารยชน
อารยชาติ
อารยธรรม
อารยประเทศ
อารัมภ
อารัมภกถา
อารัมภบท
อาวัล
อาวาห
อาศิร
อาศิรพจน์
อาศิรพาท
อาศิรพิษ
อาศิรวจนะ
อาศิรวาท
อาศิรวิษ
อาศีรพิษ
อาศีรวิษ
อาสาฬห
อาสาฬหบูชา
อิฏฐ
อิทธิปาฏิหาริย์
อินทร
อินทรธนู
อินทรวงศ์
อินทรวิเชียร
อินทรศักดิ์
อินทรายุธ
อินทรีย
อินทรียโคจร
อินทรียญาณ
อินทรียสังวร
อิ่มอกอิ่มใจ
อิสร
อิสรภาพ
อิสริย
อิสริยยศ
อี๊
อีหน็องอีแหน็ง
อีเหละเขละขละ
อึงอล
อุคห
อุคหนิมิต
อุณห
อุตตรผลคุนี
อุตตรภัทรบท
อุตตรอาษาฒ
อุตตราสาฬหะ
อุตตะภัททะ
อุตร
อุตรกุรุ
อุตรกุรู
อุตรนิกาย
อุตรผลคุนี
อุตรภัทรบท
อุตริมนุสธรรม
อุตสาห
อุตสาหกรรม
อุตสาหกรรมศิลป์
อุท

อุทบาตร
อุทบาน
อุทพินทุ์
อุบอิบ
อุปปาติก
อุปสัมปทาเปกข์
อุปสัมปทาเปกษ์
อุปัชฌาย
อุปัชฌายวัตร
อุปัทว
อุปาณ
อุ๊ย
อุยหน่า
อุ่ยหน่า
อุร
อุรค
อุรเคนทร์
อุลปนะ
อุษณ
อุษณกร
อุษณกาล
อุษณรัศมี
อุษณรูจี
อุษณีษ์
อุษาโยค
อูย
เอกภักดิ์
เอกส
เอกอัครราชทูต
เอกอัครสมณทูต
เอมโอช
เออน่ะ
เอ้อเร้อเอ้อเต่อ
เอาอกเอาใจ
แอนติเจน
แอนติบอดี
แอนติอิเล็กตรอน
โอฆ
โอฆชล
โอฆสงสาร
โอ๊ต
โอร
โอษฐ
โอษฐชะ
โอษฐภัย
โอห์มมิเตอร์
ไอ้
ไอศุริย
ไอศุริยสมบัติ
ฮวบฮาบ
ฮะฮ้าย
ฮัจญ์
ฮัจญะฮ์
ฮัจญี
ฮึดฮัด
ฮื่ย
ฮุยเลฮุย
ฮู้
เฮ็ด
แฮก
ไฮโดรเจนเพอร์-
ออกไซด์

ตารางรายการคำที่ MICROSOFT WORD 7.0 ไม่สามารถตัดแบ่งคำได้

กฎบัตรสหประ-
ชาชาติ
กฎยุทธวินัย
กบฑูต
กปณก
กรณีย
กรรมการิณี
กรรมชรูป
กรรมสัมปาทิก
กรรมาธิการ
กระจิ๋ง
กระจิ๋ม
กระเฉง
กระโชกโฮกฮาก
กระชิกกระชี้
กระชี้
กระเซ้ากระซี้
กระดิ้ง
กระเดียม
กระแดะกระแด๋
กระติ้ง
กระเตี้ยม
กระโถนฤาษี
กระทั้น
กระแทกกระทั้น
กระแท้ม
กระปุกหลุก
กระเปาะเหลาะ
กระเปี้ย
กระแป้
กระผลาม
กระเฟียด
กระแมม
กระยาคชวาง
กระยาสังแวง
กระลาพิน
กระเษมสานต์
กระเหม่า
กระอ่วน
กระอักกระอ่วน
กรัณฑ
กริตย
กริยาวิเศษณานุ
ประโยค
กรุณาทฤคุณ
กฤด
กฤตติกา
กฤตย
กฤษฎาญชเลศ
กล้องจุลทรรศน์
กล้องตุด
กล้องโทรทรรศน์
กลับตาลปัตร
กษัตราธิราช
กษัตริยชาติ
กษีรามพุ
กะดวน
กะมง
กัณฐ
กัณฐชะ
กัณฐัส
กัณฐัสว์
กัณห
กัณหธรรม
กัณหปักษ์
กัมป่
กัมปน
กัมปนาการ
กัลยาณ
กากขยาก
กาญจน
กามราค
การิตการก
การิตวาจก
กาลโยค
กาฬวาต
กำเริบเสิบสาน
กินหู้
กี
กึกกือ
กุกกุฏ
กุกกุร
กุฏาคาร
กุมภ
กุมภการ
กุมภัณฑ
กุมภัณฑยักษ์
กุมารลพิตา
กุศโลบาย
กูฏาคาร
เก๋งพั้ง
เกลือเบสิก
เกลือยูเรต
เกศี
เกษตรและ-
สหกรณ์
แกงบวด
โกกิล
โกญจ
โกร๋นเกร๋น
ขมึงทึง
ขรรค
ขล้ง
ขลึง
ขลุกขลัก
ขลุกขลิก
ขวายขวน
ขัณฑ
ขัณฑสกร
ขัตติย
ขัตติยมานะ
ข้าวโกบ

ข้าวบิณฑ์
ข้าวปุ้น
ข้าวเภา
ข้าวฮาง
ขิด
ขี้กะโล้โท้
ขี้เข็บ
ขี้ทูด
ขี้ปะติ๋ว
ขี้แย
ขี้เหร่
ขึ้นพลับพลา
ขุนพัฒน์
ขุย
ขุยไผ่
ขุยอินทรีย์
ขู
เขม
เขมรไทรโยค
เขือม
แขยงแขงขน
ไข้ทรพิษ
คช
คชกรรม
คชนาม
คชลักษณ์
คชศาสตร์
คชสาร
คชสีห์
คณ
คต
คมน
ครรไล
ครอบครอง-ปรปักษ์
คร่ำเครอะ
คริสเตียน
ครื้นครึก
คลอแคล
คล้อแคล้
คล่องแคล่ว
คลับคล้ายคลับ
คลายเคล่ง
ควีนสิริกิติ์
คห
คหกรรมศาสตร์
คหปตานี
คัดน
คันถ
คันถธุระ
คันธมาทน์
คันบวย
คัพภ
คัพภสาลี
คัมภีร
คัมภีรภาพ
ค่ายเยาวชน
คาราคาก่า
คำนูณ
คิง
คิลาน
คิลานปัจจัย
คุมนุม
คุยห
คุยหฐาน
คุยหประเทศ
คุยหรหัสย์
คู่ปรปักษ์
เคมีอนินทรีย์
เคร่งครัด
เครดิตฟองซิเอร์
เคห
เคหศาสตร์
เคหเศรษฐศาสตร์
เคอย
แค็ตตาล็อก
แคลเซียมคาร์ไบด์
ฆ้องหุ่ย
งกเงิ่น
งงงวย
ง่วงงุน
งอมแงม
งาลั่ว
งูสวัด
เงินสเตอร์ลิง
โงนเงน
จตุตถ
จตุทิพยคันธา
จตุรงคนายก
จตุรงคประดับ
จตุรงคยมก
จตุรพิธ
จตุรพิธพร
จตุลังคบาท
จรรมขัณฑ์
จริม
จริย
จริยธรรม
จริยวัตร
จริยศาสตร์
จริยศึกษา
จองคช
จะกรัจจะกราจ
จ๊ะเอ๋
จักรวาต
จัณฑ
จัณฑวาตา
จัตวาทัณฑี
จัตุร
จัตุรพักตร์
จัตุรภุช
จัตุรมุข
จันทนา
จันทร
จันทรมณฑล
จันทรวงศ์
จันทรุปราคา
จัว
จาค
จาตุทสี

จาตุร
จาตุรงค
จาตุรงคสันนิบาต
จาตุรราชการ
จำนวนจินตภาพ
จำนวนตรรกยะ
จำนวนอตรรกยะ
จิตต
จิตตภาวนา
จิตตวิสุทธิ
จิตตานุปัสสนา
จิตรปทา
จินต
จินตกวี
จินตภาพ
จิร
จิรกาล
จุฑาธิปไตย
จุดยุทธศาสตร์
จุทสมสุรทิน
จุลทรรศน์
จุลินทรีย์
เจดีย
เจดียถาน
เจดียสถาน
เจตนารมณ์
เจติย
เจริญสมณธรรม
เจ้าขุนมูลนาย
เจ้าเซ็นเต้นต๋ำบุด
เจ้าประคู้น
เจ้าสัว
เจีย
เจื้อย
เจื้อยแจ้ว
แจ่ง
แจ้น
แจบ
แจ่ม
แจ่มแจ้ง
แจ่มใส
แจ๋วแหวว
ฉพีสติม
ฉศก
ฉัน ท
ฉั นทลักษณ์
ฉัน ทวิลาส
ฉัน ทศาสตร์
ฉิน ท
ฉิน ทฤกษ์
ฉุก เ ฉิน
ฉุก ละหุก
เฉไฉ
เฉยเมย
โฉ่งฉ่าง
ชงฆ
ชนม
ชเนตตี
ชมชัว
ชลธารก
ชลธิศ
ชว
ชอื้อ
ชอุ่ม
ชัค
ชังฆ
ชังฆวิหาร
ชาต
ชาตรูป
ชาตสระ
ชาวเล
ชีรณัคดิ
ชีว
ชีวเคมี
ชีวประวัติ
ชีววิทยา
ชุณห
ชุษณ
เชษฐ
เชษฐะ
โชติก
โชโหว
ซอกแซก
ซ่องแซ่ง
ซับซาบ
ซึมซาบ
โซเดียมซัลเฟต
โซเดียม-
ไซคลาเมต
โซเดียม-
ไฮโดรเจน-
กลูทาเมต
โซเดียม-
ไฮโดรเจนซัลเฟต
โซรมศัสตราวุธ
ฌาปน
ฌาปนกิจ
ฌาปนสถาน
ญาณทัสนะ
ฐานกรณ์
ฐาปน
ฑังส
ดนุช
ดอกดั้ว
ดังรือ
ดามพ
ดามพวรรณ
ดุ่ม
ดุริย
ดุริยางค
ดุริยางคศาสตร์
ดุริยางคศิลป์
ดูรี
เด็กเมื่อวานซืน
แดะแด๋
ตกขุย
ตกใต้เถรเทวทัต
ตจ

ตจปัญจก-
กรรมฐาน
ตจสาร
ต้มโคล้ง
ต้มเปอะ
ตรมตรอม
ตรรก
ตรรกศาสตร์
ตรังค
ตรังควชิราวดี
ตราไปรษณียากร
ตรีกันสวาต
ตรีฉินทลามกา
ตรีชาต
ตรีทิพยรส
ตรีทุรวสา
ตรีเทวตรีคันธา
ตรีปีตผล
ตรีมธุร
ตรีโลจน์
ตรีโลหก
ตรีวาตผล
ตรีสุรผล
ตรีเสมหผล
ตฤตีย
ตฤๅ
ต่อนัดต่อแนง
ตอม่อ
ตะพึดตะพือ
ตะหลุง

ตาโก้ง
ตามพ
ตามพหัตถี
ตายขุย
ตาลปัตร
ตาลปัตรบังเพลิง
ตาลปัตรฤๅษี
ตาลอ
ติก
ติงส
ติงสติม
ตืดตื๋อ
ตุ๊ตะ
ตุ้บตั้บ
ตุรงค
ตุล
ตุหรัดตุเหร่
แตรฟันฟาร์
แตรวงโยธวาทิต
ไตรกิศยา
ถนัดถนี่
ถมถืด
ถมเถ
ถวัดถวัน
ถั่วคร้า
ถ่านไฟแช็ก
ถิร
ถีน
ถึงแก่มรณภาพ
ถึงแก่อสัญกรรม

ถุล
ถูปารหบุคคล
ถูล
เถย
เถยจิต
เถยเจตนา
เถยสังวาส
เถลไถล
ไถย
ทนต
ทร
ทรกรรม
ทรงข้าวบิณฑ์
ทรงมัณฑ์
ทรชน
ทรชาติ
ทรธึก
ทรพล
ทรพิษ
ทรภิกษ์
ทรยศ
ทรยุค
ทรรศนาการ
ทรราช
ทรลักษณ์
ทรัพย
ทรัพยสิทธิ
ทฤฆ
ทลบม
ทวัตดึงสาการ

ทวาทศม
ทว่าทศี
ทวาบร
ทวิช
ทศเบญจกูล
ทศพิธราชธรรม
ทศม
ทองคำเปลว
ทองเค
ท้องอัสดงคต
ทอดหุ่ย
ทักษ
ทัณฑ
ทัณฑกรรม
ทัณฑฆาต
ทัณฑนิคม
ทัณฑวิทยา
ทัณฑสถาน
ทัต
ทันต
ทันตชะ
ทันตแพทย์
ทัศน
ทัศนวิสัย
ทัศนศาสตร์
ทัศนศึกษา
ทัศนีย
ทัศนียภาพ
ทาร
ทิชากร

ทิฐ
ทิพย
ทิพยญาณ
ทิพยเนตร
ทิพยพยาน
ทิพยมานุษ
ทิพยรส
ทิพยโสตร
ทิวส
ทีฆ
ทีฆนิกาย
ทีฆสระ
ทีรฆ
ทุกข
ทุกขนิโรธ
ทุกขลาภ
ทุกขสมุทัย
ทุกขารมณ์
ทุร
ทุรคม
ทุรภิกษ์
ทุษฐ
ทุส
ทูร
เทพย
เทพยเจ้า
เทพยดา
เทเพนทร์
เทว
เทวทัณฑ์
เทวทูต
เทวธรรม
เทวธิดา
เทวนาครี
เทวนิยม
เทวภาวะ
เทวรูป
เทวโลก
เทววาจิกะ
เทววิทยา
เทวสถาน
เทห์
เทห์ฟากฟ้า
โทรทรรศน์
โทรมศัสตราวุธ
โทษกรณ์
โทส
ธงเยาวราชน้อย
ธงเยาวราชใหญ่
ธนิษฐะ
ธนุร
ธนุรมารค
ธรรมกถึก
ธรรมนูญ
ธรรมนูญการ-
ปกครองราช-
อาณาจักร
ธรรมะธัมโม
ธรรมาภิมุข
ธราธิบดี
ธราธิป
ธเรษตรีศวร
ธัญดัจ
ธาตุโขภ
ธานินทร์
ธาราธิคุณ
ธุดงค
ธุดงควัตร
ธุดงคสมาทาน
ธุม
ธุมเกตุ
ธุมชาล
ธุมเพลิง
ธุมา
ธุร
ธุรการ
ธุว
ธุวดารา
ธุวภาค
ธุวมณฑล
ธุวยัษฎี
ธูปฤาษี
นขา
นภ
นภมณฑล
นภศูล
นย
นร
นรเทพ
นรนาถ
นรนายก
นรบดี
นรบาล
นรพยัคฆ์
นรราช
นรเศรษฐ์
นรสิงห์
นรสีห์
นฤ
นฤคหิต
นฤโทษ
นฤดม
นฤตย
นฤตยศาลา
นฤตยศาสตร์
นฤตยสถาน
นฤเทพ
นฤนาท
นฤเบศ
นฤป
นฤปนีติ
นฤปเวศม์
นฤปัตนี
นฤมล
นว
นวก
นวกภูมิ
นวกรรม
นวกรรมิก
นวการ

นวกิจ
นวคุรหะ
นวคุณ
นวเคราะห์
นวทวาร
นวนิต
นวนิยาย
นวปฏล
นวมุข
นวโลหะ
นววิธ
นวอรหาทิคุณ
นวารหาทิคุณ
นักเลง
นักษัตรเนมี
นักษัตรโยค
นักสราช
นันท
นันทปักษี
นัวเนีย
นาคสังวัจฉร
นาคินทร์
นาฏย
นามไธย
นามาภิไธย
นิคมสหกรณ์
นิคมอุตสาหกรรม
นิคห
นิคหกรรม
นิตย
นิตยทาน
นิตยภัต
นิติกรณ์
นิทรารมณ์
นิทาฆ
นิพัทธ
นิพัทธกุศล
นิยยานิก
นิรย
นิรยบาล
นิรันตร
นิศาทิ
นิสภ
นีร
นีรจร
นุงนัง
เนาว
เนาวรัตน์
แน่นนันต์
แน่นแฟ้น
บงก
บงสุกูลิก
บดีวรดา
บทภาชน์
บทภาชนีย์
บทเรศ
บทวเรศ
บมิ
บร
บรทาร
บรทารกรรม
บรรพตกีลา
บราง
บริษัทเครดิต
ฟองซิเอร์
บแรง
บแรงภักษ์
บอง
บเอ
บัฐยาพฤต
บัดแมล่ง
บาทนิเกต
บาทภัฏ
บาทรช
บานเบียง
บานโภชน์
บานีโยทก
บาร์เรล
บ้าระห่ำ
บำเรอเชอภักดิ์
บิณฑ
บิตุจฉา
บุบบิบ
บุปผ
บุปผชาติ
บุปผวิกัติ
บุพพ
บุพพาษาฆ
บุพพาสาพหะ
บุพวิเทห์
บุรพผลคุนี
บุราณทุติยิกา
บุษบ
บุษบราค
บุษบวรรษ
บุษบัน
บุษป
บุษยะ
บูชนีย
บูชนียสถาน
บูรพาษาฆ
เบญจกูล
เบญจคัพย์
เบญจม
เบญจมสุรทิน
เบญจโลหกะ
เบญจวรรณ
เบญจวรรณห้าสี
เบญจางค
เบญจางค
ประดิษฐ์
เบอะบะ
โบกขร
โบกขรพรรษ
ใบพัทธสีมา
ปฏิ
ปฏิการ
ปฏิคหิต

ปฏิญญาสากลว่า-
ด้วยสิทธิมนุษย-
ชน
ปฐมโพธิกาล
ปฐมสุรทิน
ปฐมาษาฒ
ปนเป
ปปัญจ
ปปัญจธรรม
ปโย
ปโยชนม์
ปโยธร
ปโยธรา
ปโยนิธิ
ปโยราศิ
ปร
ปรตยักษ์
ปรนัย
ปรปักษ์
ปรม
ปรโลก
ปรวาที
ประคบประหงม
ประโคนธรรพ์
ประจักษ
ประจักษนิยม
ประเจิดประเจ้อ
ประดักประเดิด
ประดิ
ประดิษฐ
ประดิษฐกรรม
ประเดี๋ยวประด๋าว
ประทุษฏจิต
ประทุษฐจิต
ประลัยวาต
ประสัยห
ปรางค์ปรา
ปริณามัคคิ
ปริย
ปริยรณ
ปริยวาท
ปริเวณ
ปรู๊ดปร๊าด
ปลวังค
ปลาม้ำ
ปวัตน
ปวัตนาการ
ปสันน
ปสันนาการ
ปสัยห
ปสัยหาการ
ปสัยหาวหาร
ปะวะหลิ่ม
ปักข
ปักขคณนา
ปักขพิพาร
ปักษ
ปักษกษัย
ปักษคณนา
ปักษคม
ปักษธร
ปักษเกท
ปักษวาหน
ปักษานดร
ปักษาวสาน
ปัจจันต
ปัจจันตคาม
ปัจจันตชนบท
ปัจจันตประเทศ
ปัจจุส
ปัจจุสกาล
ปัจเจกโพธิ
ปัจโจสกาล
ปัจยาการ
ปัญจ
ปัญจนที
ปัญจม
ปัญจวัคคีย์
ปัญจสาขา
ปัญญาส
ปัณรส
ปัณรสม
ปัณรสมสุรทิน
ปัทมราค
ปาฐ
ปาณ
ปาณก
ปาณทัณฑ์
ปาณนาศ
ปาณภูต
ปาณวินาศ
ปาณิมุกต์
ปานีย
ปาป
ปาปมุต
ปาร
ปารคู
ป่าเส็งเคร็ง
ปาหุไณย
ปิงคล
ปิงคลหัตถี
ปีตามหัยกา
ปีตามหัยยิกา
ปีหลั่น
ปีมะโว้
ปุนัพพสู
ปุพพ
ปุพพผลคุนี
ปุพพะภัททะ
ปุม
ปุ่มป่ำ
ปุรพผลคุนี
ปุรพษาฒ
ปุษยะ
ปุสสะ
ปูชนีย
ปู่ย่าปู่ยำ
เปกข์
เปิดก่า
เปต

เปตวิสัย
เปตอง
เปรมปรา
เปลว
เปศลมัธย์
เปี๊ยะ
แปรปรวน
โปกขร
โปกขรพรรษ
โป๊งโย้ง
โปฐบท
โปฐปทมาส
โปร่งเปร่ง
ไปรษณีย
ไปรษณียนิเทศ
ไปรษณียบรรณ
ไปรษณียบัตร
ไปรษณียภัณฑ์
ไปรษณียวัตถุ
ไปรษณียากร
ผมชิงเกิล
ผรณ
ผรุส
ผลคุนี
ผลคุนีบูรพมาส
ผลับ
ผลิก
ผลิกศิขรี
ผลิกศิลา
ผลุนผลัน
ผลุบผลับ
ผสมผเส
ผัสส
ผากแผ้ง
ผาลคุน
ผิน
ผินหลังให้
ผีต
ผูกพัทธสีมา
เผน
เผนธรรม
เผยิบผยาบ
เผลอไผล
เผือกโทป้าด
โผฏฐัพ
โผฏฐัพธรรม
โผฏฐัพพะ
โผย
โผอน
ฝรั่งเศส
พงศ
พจน
พญาพยาต
พยักพเยิด
พยัคฆ
พยุห
พยุหบาตร
พยุหบาตรา
พยุหยาตรา
พยุหโยธา
พยุหแสนยา
พยุหแสนยากร
พรรคกลิน
พร้อมพรัก
พระทัย
พระฤๅษี
พระหารณย์
พระหารัณย์
พฤกษ
พฤกษชาติ
พฤกษทล
พฤกษเทวดา
พฤกษราช
พฤกษศาสตร์
พฤฒ
พฤฒาจารย์
พลวง
พลังงานจลน์
พลับพลา
พักตร
พัฒน
พัทธ
พัทธสีมา
พันธ
พันธกรณี
พันธบัตร
พันธมิตร
พิจิก
พิถย
พิทย
พิทยาคาร
พิพิธภัณฑสถาน
พิมพ
พิมพาภรณ์
พิริย
พิริยพฤนท์
พิริยโยธา
พิลึกกึกกือ
พิลึกพิลั่น
พี่เบิ้ม
พีร
พุทธมามกะ
พุทธศาสนิกชน
พุ่มข้าวบิณฑ์
เพ้อเจ้อ
แพทย
แพทยศาสตร์
โพธิ
โพธิปักขิยธรรม
โพธิสมภาร
โพนเพน
ไพ่ป๊อก
ไพรสาณฑ์
ฟอดแฟด
ฟัดเฟียด
เฟอะฟะ
เฟะฟะ
ไฟแช็ก
ภย
ภราตร

ภราดรภาพ
ภราดฤ
ภราดฤภาพ
ภวังค
ภักต
ภักตกฤตย์
ภักษ
ภักษการ
ภังค
ภัณฑ
ภัณฑครรภ
ภาคเสธ
ภาว
ภาวศุทธิ
ภาวะฉุกเฉิน
ภิน
ภินท
ภินทน
ภีรุก
ภุกต
ภุกตเศษ
ภุญช
ภูธเรศวร
ภูเนตุ
ภูเบนทร์
ภูเบศ
ภูเบศวร์
ภูป
ภูมิบริมาณ
ภูรโลก
ภูวนัตตรัย
ภูวไนย
เภสัชพฤกษ-ศาสตร์
เภสัช-อุตสาหกรรม
โภชน
มณีราค
มต
มธุป
มธุร
มธุรตรัย
มธุรพจน์
มธุลีห์
มนท
มนทกานติ
มนัสดาป
มนุษย
มนุษยสัมพันธ์
มโนนุกูล
มโนภินิเวศ
มยุร
มยุรคติ
มยุรฉัตร
มยุรอาสน์
มรณ
มรณกรรม
มรณธรรม
มรณบัตร
มรณภาพ
มรณันติก
มฤคศิร
มฤคศิรมาส
มฤคศิรัส
ม้วนด้วน
มหัจฉริย
มหันต
มหิป
มอมแมม
ม่อห้อม
มะม่วงหิมพานต์
มัค
มังส
มังสวิรัติ
มัช
มัชฌ
มัชฌันติก
มัชฌันติกสมัย
มัชฌิม
มัชฌิมนิกาย
มัชฌิมบุรุษ
มัชฌิมประเทศ
มัชฌิมภูมิ
มัชฌิมยาม
มัชฌิมวัย
มัชวิรัติ
มัณฑ
มัณฑน
มัณฑนศิลป์
มัธย
มัธยฐาน
มัวซัว
มาคสิร
มาฆ
มานทะลุน
มานุษย
มานุษยวิทยา
มิคสิร
มิจฉากัมมันตะ
มิญช
มิต
มิตภาณี
มิ่ม
มิศรก
มิส
มีพห
มุขบาฐ
มุตกิด
มุบมิบ
มุรธาภิเษก
มุฮัมมัด
มูค
มูรธ
มูละ
มูสัง
มูพห
เมรย
แม่น
แม่นยำ
แมลงวันสเปน

โมกข
โมกขบริสุทธิ์
โมกษ
โมฆ
โมฆกรรม
โมฆียะ
โมฆียกรรม
โมร
โมห
ไม่กี่อัฐพส
ไม้ทัณฑฆาต
ยติภังค์
ยว
ยั้ง
ยัดเยียด
ยันตร
ยันตรกรรม
ยางมะตอย
ยาปนมัต
ยามโยค
ยี้
ยุคันต
ยุคันตวาต
ยุทธ
ยุทธการ
ยุทธนาธิการ
ยุทธภัณฑ์
ยุทธภูมิ
ยุทธโยธา
ยุทธวิธี
ยุทธวินัย
ยุทธศาสตร์
ยุทธหัตถี
ยุพ
ยุพยง
ยุพเยาว์
ยุว
ยุวชน
ยุวราช
ยุวราชา
เยภุย
เยภุยนัย
เยภุยสิกา
เยาว
เยาวชน
เยาวนะ
เยาวมาลย์
เยาวยอด
เยาวราช
เยาวเรศ
เยาวลักษณ์
เยิ่นเย้อ
โยกเยก
โยค
โยคจักษุ
โยคนิทรา
โย่งเย่ง
โยธ
โยธวาทิต
โยพนมัท
รกเรี้ยว
รงค
รงควัตถุ
รถฉุกเฉิน
รมณีย
รวนเร
รโหคต
รังสรัง
รังสีคอสมิก
รัชต
รัชต
รัฐธรรมนูญ
รัฐประศาสนนัย
รัฐประศาสน-
ศาสตร์
รัฐประศาสโน-
บาย
รัตน
รัตนโกสินทร์
รัตนบัลลังก์
รัตนสิงหาสน์
รัมก
รัมณีย
รัส
ราชกิจจานุเบกษา
ราชภัฏ
ราชโยค
ราชวโรงการ
ราชหัตถเลขา
ราชองครักษ์
รุกข
รุกขชาติ
รุกขมูล
รุกขมูลิกธุดงค์
เรียบวุธ
เรื่อยเจื้อย
แรงงานและ-
สวัสดิการสังคม
ฤๅทัย
ฤๅษี
ฤๅษีแปลงสาร
ฤๅษีผสมแล้ว
ฤๅษีเลี้ยงลิง
ลฆุโภชน์
ลหุก
ลองธรรม์
ลักษณ
ลักษณนาม
ลักษณาการ
ลัดน
ลัยคต
ลายสือ
ลุ่นตุ้น
ลุ่นโตง
ลูกมโหตร
เล่นพิเรนทร์
เลป
เลอะเทอะ
เละเทะ
เลิ่กลั่ก

เลื้อยเจื้อย
โลกาภิวัตน์
โลห
โลหกุมภี
ไล่เลี่ย
ฦา
ฦาชา
ฦาสาย
วงศ
วชิร
วชิรปาณี
วชิรหัตถ์
วณ
วณบัตร
วณพันธน์
วต
วทานิย
วธ
วธก
วร
วรดนู
วรท
วรทาน
วรมหาวิหาร
วรรณ
วรรณคดี
วรรณยุต
วรวิหาร
วสันต
วสันตวิษุวัต
วัคคุวัท
วัจ
วัจกุฎี
วัจมรรค
วัชร
วัชรธาตุมณฑล
วัชรปาณี
วัชรยาน
วัชรอาสน์
วัฏ
วัฏจักร
วัฏทุกข์
วัฒน
วัฒนธรรม
วันทย
วันทยหัตถ์
วัส
วัสคณนา
ว้าก
วากจิรพัสตร์
วากย
วากยสัมพันธ์
วาต
วาตปานะ
วาตภัย
ว่าแต่เขาอิเหนา-
เป็นเอง
วาทย
วาทยกร
วานชืน
วาลธิ
วิกาลโภชน์
วิ่งระแบง
วิทย
วิทยฐานะ
วิทยาคาร
วิทยาศาสตร์-
ชีวภาพ
วินย
วิร
วิโรธน์
วิลย
วิไลวรรณ
วิวัฒน
วิวาห
วิศาข
วิศาขบูชา
วิเศษณ
วิเศษณการก
วิสย
วิสาข
วิสาขบูชา
วิสามานยนาม
วีต
เวมัติก
เวียง
ศตปที
ศตสังวัตสร์
ศระวณ
ศราทธ
ศราทธพรต
ศศ
ศศธร
ศศพินทุ์
ศศลักษณ์
ศศิขัณฑ์
ศักย
ศักยภาพ
ศักร
ศักรภพน์
ศัพท
ศัพทมูลวิทยา
ศัลย
ศัลยกรรม
ศัลยแพทย์
ศัลยศาสตร์
ศัสตร
ศัสตรกรรม
ศัสตรการ
ศัสตรศาสตร์
ศากย
ศากยเกตุ
ศากยพุทธ
ศากยมุนี
ศานต
ศานตรส
ศานติก
ศาปมุกติ์
ศาลคดีเด็กและ-
เยาวชน

ศาลเยาวชนและ-
ครอบครัว
ศาสตร
ศาสน
ศาสนจักร
ศาสนีย
ศิร
ศิรประภา
ศิลป
ศิลปกร
ศิลปกรรม
ศิลปการ
ศิลปกิจ
ศิลปวิทยา
ศิลปศาสตร์
ศิว
ศิวโมกข์
ศิวลึงค์
ศิวเวท
ศิษฏ์
ศีต
ศีตกาล
ศึกษาธิการ
ศึกษานิเทศก์
ศุกร
ศุกรวรรณ
ศุกล
ศุภ
ศุภเคราะห์
ศุภนิมิต
ศุภมัสดุ
ศุภมาตรา
ศุภร
ศุภอักษร
ศุลก
ศูนย
ศูนยภาพ
ศูนย์เยาวชน
ศูนยวาท
เศรษฐ
เศรษฐศาสตร์
เศาจ
ษมายุมแปลง
สกฏ
สกฏภาร
สกลมหาสังฆ
ปริณายก
สงสารวัฏ
สตัมภ์
สถวิร
สถวีร
สนุกเกอร์
สภาว
สภาวการณ์
สมณ
สมณบริขาร
สมณโวหาร
สมณศักดิ์
สมณสารูป
สมณสาสน์
สมถ
สมถยานิก
สมถวิปัสสนา
สมบัตินักเลง
สมปฤๅดี
สมฤๅดี
สมุห
สมุหกลาโหม
สมุหเทศาภิบาล
สมุหนาม
สมุหนายก
สมุหประธาน
สมุหพระกลาโหม
สรณ
สรณคมน์
สรณตรัย
สรีร
สรีรกิจ
สรีรธาตุ
สรีรวิทยา
สรีรศาสตร์
สลบไสล
สลาง
สว
สวกรรม
สวการย์
สวนีย
สวภาพ
สวรรค
สวราชย์
สวัสดิการ
สวัสดิ
สวามินี
สห
สหกรณ์
สหการ
สหชาต
สหธรรม
สหธรรมิก
สหบาน
สหประชาชาติ
สหพันธ์
สหพันธรัฐ
สหภาพ
สหภาพแรงงาน
สหศึกษา
สหัส
สหัสธารา
สหัสเนตร
สะบักสะบอม
สะบัดสะบิ้ง
สังข
สังขกร
สังขต
สังขตธรรม
สังค
สังคญาติ
สังฆ
สังฆกรรม
สังฆการี

สังฆเถระ
สังฆนายก
สังฆเภท
สังฆมณฑล
สังฆมนตรี
สังฆสภา
สังฆาธิการ
สัชฌ
สัชฌกร
สัชฌการ
สัญชาต
สัญชาตญาณ
สัตตาห
สัตม
สัตมวาร
สัตย
สัตยพรต
สัตว
สัตวบาล
สัตวแพทย์
สัตวโลก
สัตววิทยา
สัทธาธิก
สันถว
สันถวไมตรี
สันทิฐิก
สับปลี้
สัปด
สัปดปกรณ์
สัปดประกรณ์
สัปดสดก
สัปต
สัปปุริส
สัมพัจฉร
สัมพัจฉรฉินท์
สัมพันธ
สัมพันธภาพ
สัมพันธมิตร
สัมพันธไมตรี
สัมพาหะ
สัมพุทธ
สัมฤทธิ
สัมฤทธิศก
สัลลาป
สัลเลข
สัสต
สัสตทิฐิ
สากษิน
สากหยาก
สากิย
สากิยบุตร
สากิยมุนี
สาง
สาฏิก
สาณ
สาธย
สามนต
สามนตราช
สามานย
สามานยนาม
สารบบ
สารบับ
สารัตถ
สาราณีย
สาราณียธรรม
สาหัตถ
สาหัตถิก
สำนึง
สิขานล
สิงห
สิงหนาท
สิงหบัญชร
สิงหราช
สิงหลก
สิญจ
สิทธ
สิร
สีฆ
สีต
สีห
สีหนาท
สีหบัญชร
สีหราช
สุคนธ
สุคนธรส
สุญญ
สุทธ
สุนทรีย
สุพรรณภาชน์
สุภ
สุ้มเสียง
สุร
สุรคต
สุรบถ
สุรภาพ
สุรโลก
สุรสีหนาท
สุรเสียง
สุริย
สุริยกันต์
สุริยกานต์
สุริยการ
สุริยกาล
สุริยคติ
สุริยคราส
สุริยเคราะห์
สุริยมณฑล
สูป
สูรย
สูรยกานต์
สูรยคราส
เสกข
เสกขบุคคล
เสข
เสขบุคคล
เสร็งเคร็ง
เสต
เสนห
เสนาธิการ
เสนาธิปัต

เสนาพยุห์
เสนาพยูห์
เสล
เสลบรรพต
เสารภย์
เสาว
เสาวคันธ์
เสาวธาร
เสาวภา
เสาวภาพ
เสาวรภย์
เสือบอง
แสนย
โสตทัศนวัสดุ
โสตทัศนอุปกรณ์
โสภ
หงส
หงสคติ
หงสบาท
หงสรถ
หงสราช
หงำเหงอะ
หญ้ายายเภา
หทย
หน็องแหน็ง
หนอนตายหยาก
หมักหมม
หมากฮอส
หม้ำ
หม้ำตับ
หมุบหมิบ
หมูหย็อง
หย
หยอกเอิน
หย็อง
หย็องกรอด
หย็อมแหย็ม
หย็อย
หยิกหย็อง
หรณ
หรรษ
หรัสว
หรัสวมูรดี
หฤทย์
หฤทัยกัปน์
หฤทัยกัมป์
หฤทัยปรีย์
หฤทัยพันธน์
หลงใหล
หลีบ
หลู่หลี่
หว่าง
หสน
หอมเตียม
หังส
หัตถการ
หัตถกิจ
หัตถบาส
หัตถพันธ์
หัตถศึกษา
หัตถาภรณ์
หัวหน่าว
หัสต
หิม
หิมพาน
หิมพานต์
หิมวัต
หิมวันต์
หิมวา
หิมวาต
หิมวาน
หิมวาส
หิมเวศ
หิรัณย
หิรัณยเกศ
หิรัณยรัศมี
หีน
หีนยาน
หึงส
หื่นหรรษ์
หือรือโหด
หุต
หุตโกวิท
หู่หี่
เหมันต
เหมาะเหม็ง
แหย็ม
อกัปปิย
อกัปปิยวัตถุ
อกัปปิยโวหาร
อคาร
องค
องค์การสห
ประชาชาติ
องคชาต
องครักษ์
องควิการ
องควิเกษป
องควิทยา
อจล
อจิร
อดุลย
อทระ
อธิก
อธิกวาร
อธิกสุรทิน
อน
อนงค
อนงคเลข
อนงคเลขา
อนธ
อนัตถ
อนันต
อนันตร
อนากูล
อนารย
อนารยชน
อนารยธรรม
อนึก
อนึกทรรศนะ

อนีกทัศนะ
อนุตร
อนุรักษ
อนุรักษนิยม
อนุราธ
อนุราธะ
อเนกอนันตัง
อปจายน
อปจายนธรรม
อปจายนมัย
อปร
อปรภาค
อปรัณณชาติ
อปาน
อภิชนาธิปไตย
อภิณห
อมต
อมตบท
อย
อรทัย
อรรถกถาจารย์
อรหันต
อรหันตฆาต
อริย
อริยผล
อวิญญาณก
อวิญญาณกรัพย์
อวิโรธน์
อสังหาริม
อสัญ
อสัญกรรม
อสัมภิน
อสัมภินพงศ์
อสัมภินวงศ์
อสิเลสะ
อเสกข
อเสกขบุคคล
อเสข
อเสขะ
อ้อยส้อย
อะดรีนาลีน
อะดักอะเดี้ย
อะดักอะเดื่อ
อะดักอะแด้
อักขร
อักขรวิธี
อักขรวิบัติ
อัคนิพ่าห์
อัคนิโหตร
อัครสมณทูต
อังส
อังสกุฏ
อัจฉริย
อัจฉริยบุคคล
อัจฉริยภาพ
อัจฉริยลักษณ์
อัจฉริยลักษณะ
อัชฌัตติก
อัญ
อัญดิตถีย์
อัญเดียรถีย์
อัญมณี
อัญมัญ
อัฏฐ
อัฐม
อัฐพส
อัฐมิญชะ
อัฒ
อัฒจันทร์
อัฒมณฑล
อัฒมาส
อัฒรัตติ
อัณฑ
อัณฑโกส
อัณฑชะ
อัต
อัตชีวประวัติ
อัตนัย
อัตภาพ
อัตวินิบาตกรรม
อัตหิต
อัตเหตุ
อัทธ
อัทธคต
อัทธคู
อันต
อันตกิริยา
อันตคู
อันตลิกขะ
อั้นตั้น
อันติม
อันติมสัจ
อันเตวาสิก
อันธ
อันโยนย
อันวย
อัป
อัปเปหิ
อัพภันตร
อัมพ
อัมพวัน
อัมพวา
อัศวมุขี
อัศวเมธ
อัศวยุช
อัษฎ
อัษฎกฉัน ท์
อัษฎมงคล
อัษฎมงคล
อัส
อัสดงคต
อัสมุขี
อัสสนี
อาคันตุก
อาคันตุกภัต
อาคันตุกวัตร
อาจริย
อาจริยวาท
อาชีว
อาชีวศึกษา

อาทิตย
อาทิตยมณฑล
อาทิตยวาร
อานาปานัสสติ
อ้ายเจี๊ยว
อายุร
อายุรแพทย์
อายุรเวช
อายุรศาสตร์
อารย
อารยชน
อารยชาติ
อารยธรรม
อารยประเทศ
อารัมภ
อารัมภกถา
อารัมภบท
อาวัล
อาวาห
อาศิร
อาศิรพจน์
อาศิรพาท
อาศิรพิษ
อาศิรวจนะ
อาศิรวาท
อาศิรวิษ
อาศีรพิษ
อาศีรวิษ
อาสาฬห
อาสาฬหบูชา
อิฏฐ
อินทร
อินทรธนู
อินทรวงศ์
อินทรวิเชียร
อินทรศักดิ์
อินทรีย
อินทรียโคจร
อินทรียญาณ
อินทรียสังวร
อิสร
อิสรภาพ
อิสริย
อิสริยยศ
อีหน็องอีแหน็ง
อีเหละเขละขละ
อึงอล
อุคห
อุคหนิมิต
อุณห
อุตตรผลคุนี
อุตตรภัทรบท
อุตตรอาษาฒ
อุตตราสาฬหะ
อุตตะภัททะ
อุตร
อุตรกุรุ
อุตรกุรู
อุตรนิกาย
อุตรผลคุนี
อุตรภัทรบท
อุตริมนุสธรรม
อุตสาห
อุตสาหกรรม
อุตสาหกรรมศิลป์
อุท
อุทบาตร
อุทบาน
อุทพินท์
อุบอิบ
อุปปาติก
อุปสัมปทาเปกข์
อุปสัมปทาเปกษ์
อุปัชฌาย
อุปัชฌายวัตร
อุปัทว
อุปาณ
อุ๊ย
อุยหน่า
อุ่ยหน่า
อุร
อุรค
อุรเคนทร์
อุษณ
อุษณกร
อุษณกาล
อุษณรัศมี
อุษณรูจี
อุษาโยค
อูย
เอกภักดิ์
เอกส
เอกอัครสมณทูต
เอมโอช
เออน่ะ
เอ้อเร้อเอ้อเต่อ
แอนติอิเล็กตรอน
โอฆ
โอฆชล
โอ๊ต
โอร
โอษฐ
โอษฐภัย
ไอ้
ไอศุริย
ไอศุริยสมบัติ
ฮวบฮาบ
ฮะฮ้าย
ฮัจญ์
ฮัจญะฮ์
ฮัจญี
ฮึดฮัด
ฮี่ย
ฮุยเลฮุย
ฮู้
เฮ็ด
แฮก

ภาคผนวก ค

การตัดพยางค์คำไทย โดยโครงสร้างข้อมูลแบบสแตกและดิกชันนารีแบบแฮชชิ่ง

Thai Word Separator using Double Stack with Hashing Dictionary

สรศักดิ์ ไทยแท้
นักศึกษาปริญญาโท คณะเทคโนโลยีสารสนเทศ

บุญธีร์ เครือตราชู
อาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

เนื่องจากตัวอักษรไทยมักจะเขียนติดกันไม่มีตัวอักษรพิเศษขั้นระหว่างคำดังในภาษาอังกฤษ จึงเป็นการยากที่จะทราบว่าผู้เขียนต้องการจะแบ่งคำอย่างไร หรือ คำใดมีคำที่สะกดผิดอยู่ ทั้งนี้การรู้ขอบเขตของคำโดยอัตโนมัติจะช่วยให้สามารถตรวจสอบคำสะกดได้ถูกต้องและรวดเร็วขึ้นทั้งสามารถทำ index จากคำในข้อความได้ งานวิจัยนี้พยามตัดแบ่งคำโดยอัตโนมัติโดยใช้พจนานุกรม (Dictionary) แบบแฮชชิ่ง (Hashing) เป็นตัวช่วยในการตัดคำ รวมทั้งได้ออกแบบ Algorithm แบบ Loop และโครงสร้างข้อมูล Stack ที่ใช้ ในการเก็บตำแหน่งคำในการตัดคำแทนแบบเดิมที่เป็น Recursive รวมทั้งพจนานุกรมที่ใช้แบบ Hash ทำให้สามารถเพิ่มเติมคำได้ง่ายรวมทั้งสามารถเลือกหาคำคล้ายได้สะดวก งานวิจัยนี้ยังได้ทำการเปรียบเทียบการตัดคำและการตรวจสอบคำผิดกับ Software คือ Word 6.0 และ Word 7.0 รวมทั้งได้รวบรวม ข้อผิดพลาดของ Word 6.0 และ Word 7.0 มาทำการปรับปรุง Algorithm ในงานวิจัยนี้

Abstract

Thai alphabets are written adjacent, without any special alphabet between words. Therefore, it is difficult to separate each word in a line, or to detect any misspelling word. Separate words automatically will help both spelling correction and also document indexing. This research present the new Thai word separater using Hashing dictionary, and design the new separator algorithm using Loop instead of Recursive with double stack data structure. We can also make word suggestion to the incorrect word with little modification to the dictionary. Hashing dictionary is also easier to add new words. The new algorithm and dictionary has been compared with standard softwares Word 6.0 and Word 7.0 in both the running time and correctness.

1. บทนำ

ในปัจจุบันได้มีการใช้งานเวิร์ดโปรเซสซิ่งภาษาไทยกันอย่างแพร่หลาย โปรแกรมการตัดคำและการตรวจสอบตัวสะกดในงานเอกสารจึงมีความจำเป็นอย่างมาก จึงได้มีผู้วิจัยผลิตผลงานออกมากันอย่างต่อเนื่อง ซึ่งในแต่ละผลงานจะมีจุดมุ่งหมายให้การตรวจสอบคำสะกดหรือการตัดคำที่มีความถูกต้องมากที่สุดในเวลาที่รวดเร็ว และยังสามารถค้นหาคำคล้ายคำที่ผิดได้อย่างถูกต้อง

2. หลักการและลักษณะของภาษาไทย

จากหลักการของพจนานุกรมสามารถแบ่งอักษรไทยได้ 3 กลุ่ม ดังนี้

- กลุ่มที่ 1 ได้แก่ กลุ่มวรรณยุกต์ ประกอบด้วย ไม้ยมก ไม้เอก ไม้โท ไม้ตรี ไม้จัตวา ไม้ไต่คู้ การันต์
- กลุ่มที่ 2 ได้แก่ กลุ่มสระ ประกอบด้วย อะ อา อิ อี อึ อื อำ อุ อู เอ แอ ไอ ใอ โอ ไม้หันอากาศ
- กลุ่มที่ 3 ได้แก่ กลุ่มพยัญชนะทั้ง 44 ตัว

ในการเขียนประโยคในภาษาไทยนั้นจะเป็นการเขียนที่ติดกันตามลำดับ การแบ่งคำจะไม่มีกฎเกณฑ์ที่แน่นอนในการตัดแบ่งคำเหมือนในภาษาอังกฤษ โดยการแบ่งคำจะขึ้นอยู่กับแต่ละคนที่ต้องการจะสื่อความหมาย ดังนั้นการที่จะทำให้คอมพิวเตอร์ทราบถึงการตรวจสอบคำนั้น คอมพิวเตอร์จะต้องทราบถึงจุดสิ้นสุดของคำเสียก่อน แต่มีในบางประโยคที่มีคำที่กำกวมยากต่อการตัดแบ่งคำ เช่น

เธอมารอกลับบ้าน

ในการตัดแบ่งคำนั้นจะสามารถทำได้เพียงการตรวจสอบว่าคำนั้นมีในพจนานุกรมหรือไม่ โดยจะสามารถแบ่งคำได้ดังนี้

เธอ * มา * รอ * กลับ * บ้าน

เธอ * มาร * อก * ลับ * บ้าน

เธอ * มา * รอก * ลับ * บ้าน

ซึ่งจะเห็นว่า "มารอกลับ" สามารถตัดแบ่งได้หลายกรณี

2.1 สรุปผลงานวิจัยเกี่ยวกับการแบ่งแยกคำในภาษาไทย

สามารถแบ่งได้ 2 แนวทาง

1. การตรวจสอบทางอักษรวิธี จะทำการตรวจสอบการสะกดคำโดยการตรวจสอบพยัญชนะ[5], สระ, วรรณยุกต์, การันต์ ว่าเป็นไปตามหลักภาษา"หรือไม่ เช่น

พยัญชนะ - ฌ ญ ฎ ฒ ฌ ฬ ไม่ปรากฏเป็นอักขระต้นของคำนอกจากจะมีพวกสระนำหน้ามาก่อน ยกเว้น ฌ

สระ - สระหลัง (า, ะ เป็นต้น) จะต้องมีพยัญชนะหรือวรรณยุกต์นำหน้าเท่านั้น

- สระหน้า (เ, แ, โ, ไ, ใ เป็นต้น) จะต้องตามด้วยพยัญชนะ

วรรณยุกต์ - เป็นอักขระแรกไม่ได้

- ต้องตามด้วยตัวสะกดหรือสระหลัง

การันต์ - ต้องเป็นตัวสุดท้ายของคำเสมอ ยกเว้นคำที่มาจากภาษาอื่น

ข้อดีของการตรวจสอบทางอักษรวิธีคือสามารถตรวจสอบได้อย่างรวดเร็ว แต่มี ข้อเสีย ในความแม่นยำ ถูกต้อง ของการตรวจสอบจะน้อย เพราะอาจจะพบคำที่ผิดแต่เป็นไปตามกฎ

2. การตรวจสอบกับพจนานุกรม จะเป็นการนำคำในประโยคมาเปรียบเทียบกับคำในพจนานุกรม โดยการตรวจวิธีนี้ได้มีการจัดเก็บคำในพจนานุกรมได้ 3 แบบ [7,8] คือ

1. แบบ Table-Table-Linear-Search
2. แบบ Table-Index-Search
3. แบบ Tree Structure [6]

ซึ่งมีข้อดี คือ มีความแม่นยำ ถูกต้องสูงแต่ ความเร็วในการประมวลผลจะช้ากว่าแบบแรก

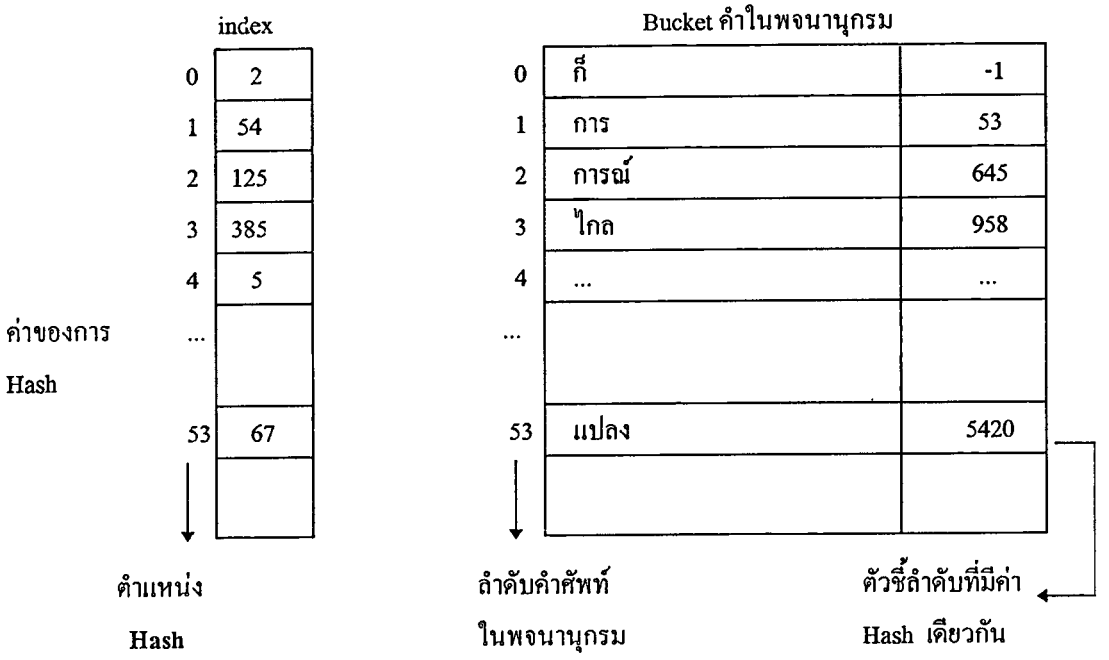
ลักษณะของคำที่เกิดในงานเอกสาร [1,2]

1. การพิมพ์เกิน เช่น การ<u>ตตรวจ</u>สอบตัวสะกด
2. การพิมพ์ผิด เช่น การ<u>ครวจ</u>สอบตัวสะกด
3. การพิมพ์ขาด เช่น การ<u>รวจ</u>สอบตัวสะกด
4. การพิมพ์ที่มีการสลับที่ระหว่างตัวอักษรที่ถูกในคำ เช่น การ<u>รตวจ</u>สอบตัวสะกด

2.2 โครงสร้างข้อมูลในพจนานุกรม

ในงานวิจัยนี้ได้ทำการเก็บคำในพจนานุกรมโดยการที่นำคำในพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 แต่ละคำมาทำการคำนวณผ่านค่าของ Hash Function เพื่อ

จะได้นำค่า Address มาทำการเก็บเป็นโครงสร้าง [3,4] ดัง รูปที่ 1



รูปที่ 1 โครงสร้างพจนานุกรมแบบ Hashing

ซึ่งมีข้อดี คือ สามารถค้นหาคำได้อย่างรวดเร็ว สามารถเพิ่มเติมคำใหม่ในพจนานุกรมได้ง่าย แต่มักจะมีการชนกันของ Address ซึ่งจะขึ้นกับค่าของ Hash Function

3. การทดลองและผลงานที่ได้จากงานวิจัย

ผลของการเก็บโครงสร้างของพจนานุกรมแบบ Hashing

ตารางที่ 1 จำนวนการเกิด Collision ของพจนานุกรม

จำนวน Collision	จำนวนคำ
ไม่มี Collision	20540
1 Collision	8484
2 Collision	2566
3 Collision	635
4 Collision	134
5 Collision	19
6 Collision	3
รวม	32,381
ค่าเฉลี่ย	0.50

การเก็บโครงสร้างของพจนานุกรมแบบ Hashing นั้นจะพบว่ามีการชนกันของ address ในการคำนวณผ่าน Hash Function ดังนี้ จากจำนวนคำทั้งหมดในพจนานุกรม = 32381 คำ พบค่าที่เกิด collision มากที่สุด 6 collision มี 3 คำโดยจะมีค่าเฉลี่ยของการเกิด collision ในพจนานุกรม 1.50 คำต่อครั้ง

ตารางที่ 2 ตารางแสดงข้อมูลในพจนานุกรม

ข้อมูล	
จำนวนไบต์ของพจนานุกรม	283,903
จำนวนคำทั้งหมด	32,381
ค่าเฉลี่ยจำนวนตัวอักษรต่อคำ	6.76

ตารางที่ 3 ตารางแสดงการแจกแจงความถี่ตัวอักษรในพจนานุกรม

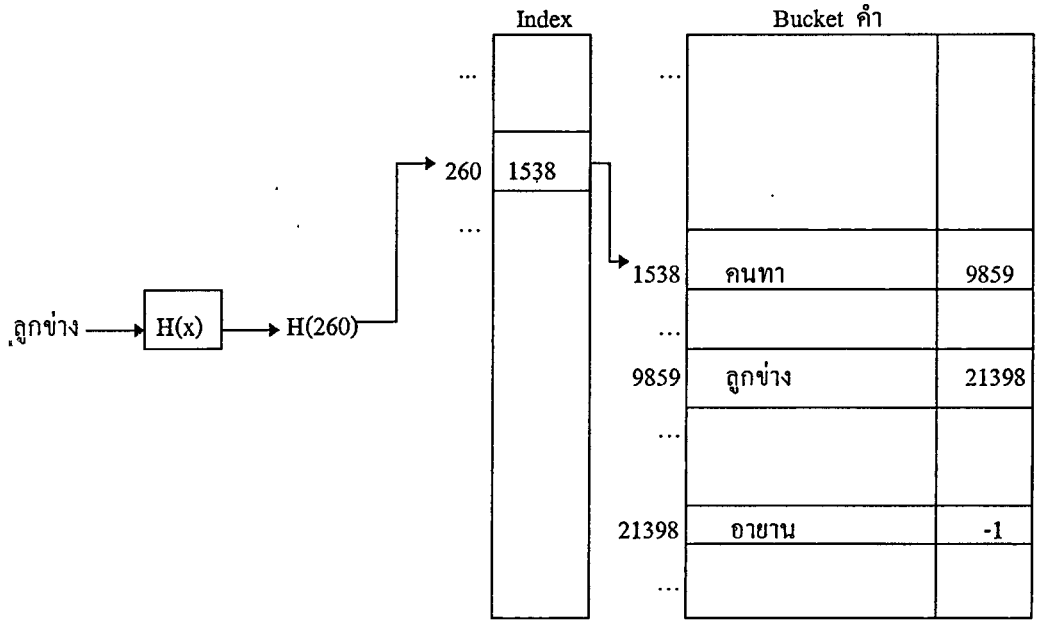
ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์ %	ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์ %	ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์ %
1	1	0.003	14	336	1.038	27	11	0.033
2	471	1.455	15	215	0.664	28	3	0.009
3	2522	7.789	16	150	0.463	29	2	0.006
4	4031	12.449	17	90	0.278	30	1	0.003
5	4890	15.101	18	77	0.238	31	2	0.006
6	5597	17.285	19	49	0.151	32	3	0.009
7	4660	14.391	20	38	0.117	33	2	0.006
8	3408	10.525	21	41	0.127	34	2	0.006
9	2177	6.723	22	18	0.056	35	1	0.003
10	1489	4.598	23	10	0.031	36	1	0.003
11	934	2.884	24	10	0.031	37	0	0
12	695	2.146	25	9	0.028	38	1	0.003
13	426	1.316	26	8	0.025	39	0	0

ตารางที่ 4 การแจกแจงความถี่ของตัวอักษรที่พบของคำในพจนานุกรม

ก	=11052	ฑ	= 229	ม	= 7946	◌ั	= 7950	◌์	= 822
ข	= 2557	ฒ	= 59	ย	= 7098	า	= 14858	◌็	= 5431
ฃ	= 0	ณ	= 1226	ร	= 12825	◌ำ	= 1679	◌้	= 6430
ค	= 3430	ด	= 4934	ฤ	= 314	◌ิ	= 6481	◌๊	= 228
ฅ	= 0	ต	= 5626	ล	= 6517	◌ี	= 4562		
ฆ	= 177	ถ	= 1007	ฦ	= 6	◌ึ	= 623		
ง	= 7600	ท	= 3930	ว	= 6172	◌ื	= 1837		
จ	= 2879	ธ	= 1202	ศ	= 1330	◌ุ	= 4075		
ฉ	= 545	น	=11369	ษ	= 928	◌ู	= 2032		
ช	= 2446	บ	= 4237	ส	= 5470	เ	= 8006		
ซ	= 618	ป	= 3944	ห	= 4880	แ	= 2484		
ฌ	= 43	ผ	= 944	ฬ	= 104	โ	= 2122		
ญ	= 831	ฝ	= 297	อ	= 8791	ใ	= 652		
ฎ	= 133	พ	= 3464	ฮ	= 176	ไ	= 1468		
ฏ	= 216	ฟ	= 466	ฯ	= 2	ๅ	= 22		
ฐ	= 298	ภ	= 1082	ะ	= 5445	ๆ	= 130		

วิธีการค้นคำในพจนานุกรมที่มีโครงสร้างแบบ Hashing

ในการค้นหาคำแบบ Hashing Search เป็นการค้นหาที่มีความรวดเร็ว ทำให้มีประสิทธิภาพในการค้นหาเป็น O(n) โดย n เป็นจำนวนคำในพจนานุกรม สำหรับพจนานุกรมภาษาไทยจะมีประสิทธิภาพของการค้นหาคำขึ้นอยู่กับจำนวนของตัวอักษรของคำที่ประกอบเป็นประโยค โดยในการเก็บพจนานุกรมแบบ Hashing นี้จะทำให้มีค่าเฉลี่ยในการค้นหาอยู่ที่ ผลคูณของค่าเฉลี่ยของการค้นหา (collision) กับค่าเฉลี่ยของจำนวนตัวอักษรที่ประกอบเป็นหนึ่งคำ ในที่นี้ คือ 1.5 x 6.76 = 10.14 ตัวอักษร ดังนั้นในการเปรียบเทียบจะทำการเปรียบเทียบประมาณ 11 ตัวอักษร หรือมีการค้นหาประมาณ 2 ครั้งในลิงก์ต่อการค้นหาในหนึ่งคำของภาษาไทย ดังตัวอย่างเช่นจะทำการค้นหาคำว่า " ลูกข่าง " จะนำคำนั้นมาทำการคำนวณผ่าน Hashing Function เพื่อจะได้ตำแหน่ง address ของคำนั้น ในที่นี้ค่า hash สุดการคำนวณเท่ากับ 260 ซึ่งจะนำไปค้นหาในโครงสร้างพจนานุกรมแบบ Hashing ดังรูปที่ 2



รูปที่ 2 รูปแสดงวิธีการค้นหาคำในพจนานุกรมแบบ Hashing

ขั้นตอนการตัดแบ่งคำโดยใช้ข้อมูลจากพจนานุกรม

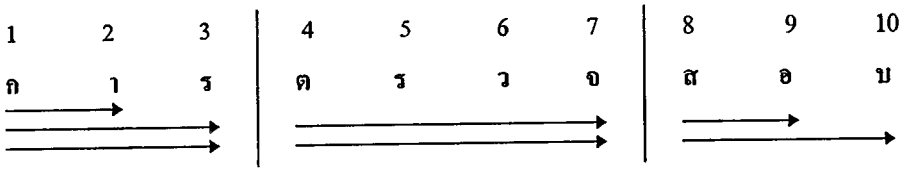
ในการตัดแบ่งพยางค์จากพจนานุกรมนี้จะใช้หลักการเปรียบเทียบกับพจนานุกรมที่มีอยู่ในหน่วยความจำ เช่น มีประโยคตัวอย่างดังนี้

1 2 3 4 5 6 7 8 9 10

ก า ร ต ร ว จ ส อ บ

ในการพิจารณาในการตัดแบ่งคำในภาษาไทย จะนำตัวอักษรของคำมาทีละตัว มาทำการคำนวณผ่าน Hash Function เพื่อทำการเปรียบเทียบกับโครงสร้างของพจนานุกรม ว่าพบคำหรือไม่พบในพจนานุกรม ถ้าพบจะนำค่าตำแหน่งของตัวอักษรมาเก็บใน Stack ตัวที่ 1 และเมื่อพบตำแหน่งที่ยาวที่สุดในการตัดคำในแต่ละรอบจะนำมาเก็บไว้ใน Stack ตัวที่ 2 แล้วเริ่มต้นคำใหม่ เริ่มจากตำแหน่งของคำที่ยาวที่สุดทำจนสิ้นสุดบรรทัด ดังตัวอย่าง คือการนำเอาตัวอักษรแรก คือ ก มาทำการคำนวณผ่าน Hash Function เพื่อเปรียบเทียบโครงสร้างข้อมูลในพจนานุกรมที่เก็บแบบ Hashing ว่าพบหรือไม่ ถ้าพบจะทำการใส่ค่าตำแหน่งตัวอักษรใน Stack ตัวที่ 1 และถ้าไม่พบจะทำการนำตัวอักษรต่อไป คือ า มาทำการคำนวณต่อจากตัวอักษรก่อนหน้า แล้วนำมาเปรียบเทียบกับพจนานุกรมว่าพบหรือไม่ จะทำเช่นนี้จนกว่าจะสิ้นสุด

บรรทัด และเมื่อได้คำที่ยาวที่สุดจะทำการเก็บตำแหน่งไว้ ใน Stack ตัวที่ 2 เพื่อให้ทราบถึงตำแหน่งที่ตัดคำดังรูปที่ 3



10	(สอบ)
9	(สอ)
7	(ตรวจ)
3	(การ)
2	(กา)

ค่าใน Stack ตัวที่ 1
จะเก็บตำแหน่งที่สามารถตัดได้

10	(สอบ)
7	(ตรวจ)
3	(การ)

ค่าใน Stack ตัวที่ 2
จะเก็บตำแหน่งที่ตัดยาวที่สุดได้

รูปที่ 3 รูปแสดงค่าตำแหน่งที่สามารถทำการตัดแบ่งคำได้

การถอยหลังเมื่อพบปัญหา (Back Tracking)

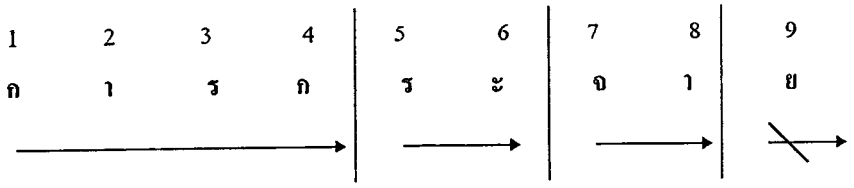
ในการค้นหาคำในพจนานุกรมเพื่อนำมาตัดแบ่งคำ อาจจะพบปัญหาในการตัดแบ่งคำบางคำ ซึ่งในงานวิจัยนี้ได้นำวิธีการ Back Tracking มาใช้เพื่อใช้ตัดสินการตัดแบ่งคำ เพื่อให้ได้คำที่ถูกต้องมากที่สุด ดังตัวอย่าง

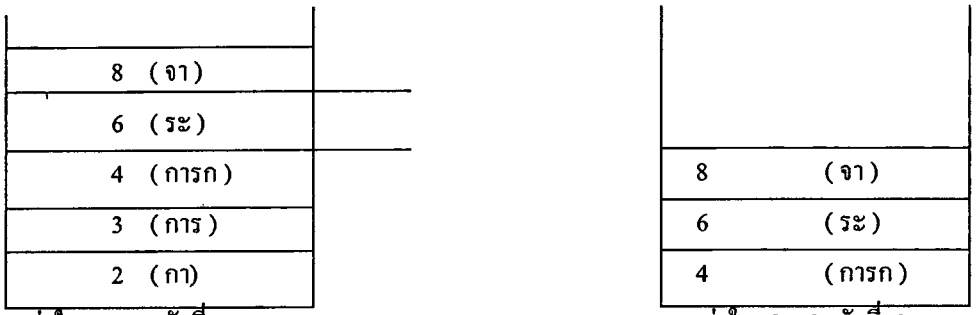
1 2 3 4 5 6 7 8 9

ก า ร ก ร ะ จ า ย

ในการทำงานจะมี ขั้นตอนดังนี้ คือ เริ่มต้นจะหาคำที่ยาวที่สุดก่อน ในที่นี้จะได้คำว่า "การก" และจะทำการเก็บตำแหน่งในแต่ละตัวใน stack ตัวที่ 1 ว่ามีคำที่สั้นกว่าอะไรบ้าง คือคำว่า กา, การ จากนั้นในรอบต่อมานำตัวอักษรถัดจากคำว่า การก มาตัดแบ่งต่อจะได้คำเดียวคือ "ระ" และนำตัวอักษรที่ต่อจาก ระ มาตัดแบ่งต่อจะได้คำว่า จา แล้วนำตัวอักษรถัดมาตัดแบ่งโดยในที่นี้จะมีตัวอักษร ย เหลือเพียงตัวเดียว ซึ่งไม่ สามารถตัดแบ่งได้ต่อ จึงทำการ Back Track ไปยังตำแหน่งก่อนหน้า รวมทั้งทำการเก็บตำแหน่งของคำที่ยาวที่สุดในการตัดแบ่งคำไว้ด้วย โดยไปยังที่ตัวอักษร จ (ตำแหน่งที่ 7) แล้วทำการคำนวณผ่าน Hash Function ถ้าพบคำในพจนานุกรมจะทำการเก็บค่าใน Stack ทั้ง 2 ตัวใหม่ โดยการ Back Track จะทำไปจนกว่าค่าที่ Back Track ใน Stack ตัวที่ 1 จะหมด ซึ่งจะมีขั้นตอนการทำงานดังนี้

ขั้นตอนที่ 1

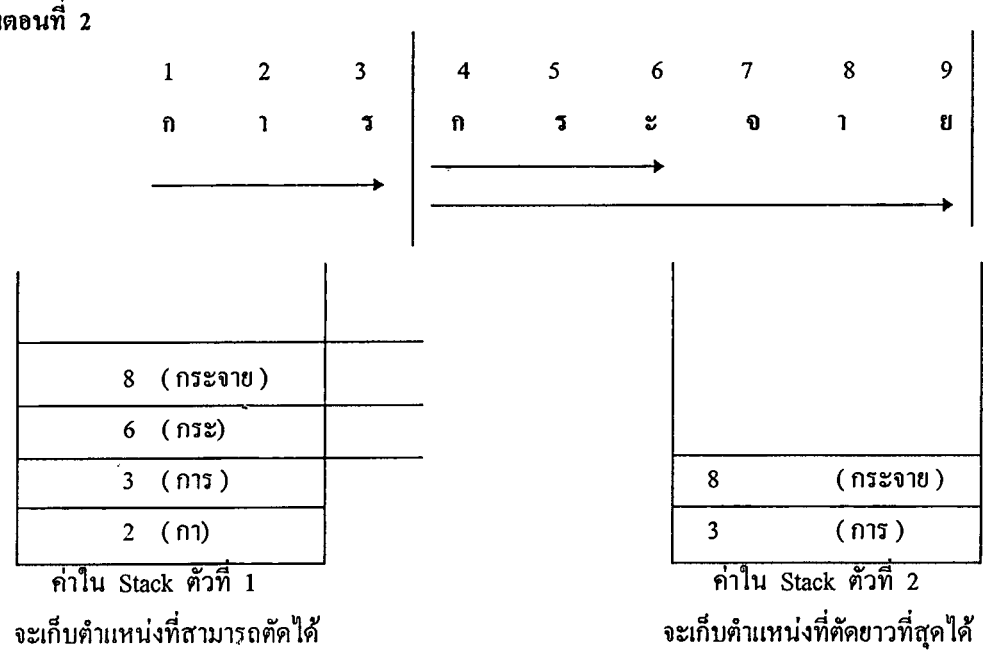




ค่าใน Stack ตัวที่ 1 จะเก็บตำแหน่งที่สามารถตัดได้

ค่าใน Stack ตัวที่ 2 จะเก็บตำแหน่งที่ตัดยาวที่สุดได้

ขั้นตอนที่ 2



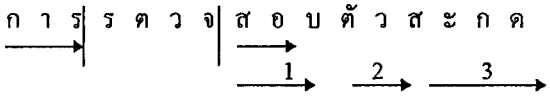
ค่าใน Stack ตัวที่ 1 จะเก็บตำแหน่งที่สามารถตัดได้

ค่าใน Stack ตัวที่ 2 จะเก็บตำแหน่งที่ตัดยาวที่สุดได้

รูปที่ 4 รูปแสดงขั้นตอนตำแหน่งการตัดคำใน Stack ในการ Back Tracking

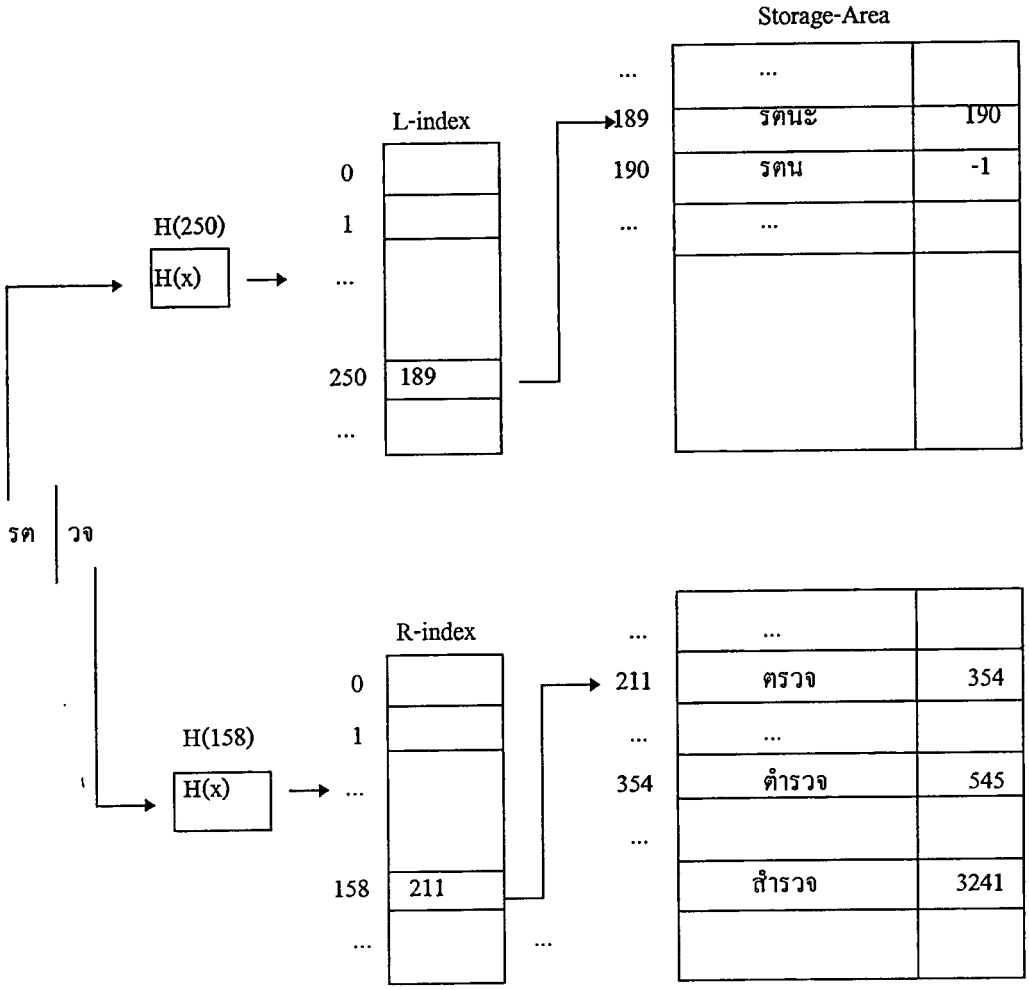
กรณีหาขอบเขตของคำที่ไม่พบในพจนานุกรม

ในการตัดแบ่งคำเปรียบเทียบกับพจนานุกรม เมื่อทำการ Back Track จนไปถึงที่ตัวอักษรต้นบรรทัด โปรแกรมจะทำการหาขอบเขตของคำที่ไม่พบในพจนานุกรม โดยจะทำการเก็บค่าตำแหน่งที่ทำการตัดไปแล้วมากที่สุดเป็นตำแหน่งเริ่มต้นของคำที่ผิด แล้วทำการ Skip ตัวอักษรไปทีละตัว เพื่อนำตัวอักษรตัวถัดไปมาทำหาคำในพจนานุกรม เมื่อพบว่าจากตำแหน่งที่คำนวณใหม่สามารถทำการตัดแบ่งคำได้เมื่อเปรียบเทียบกับพจนานุกรมได้มากกว่าหรือเท่ากับสองคำที่พบในพจนานุกรม โปรแกรมจะนำตำแหน่งนั้นมาเป็นจุดสิ้นสุดคำที่ผิด เพื่อทำการแก้ไขต่อไป ดังตัวอย่างต่อไปนี้



วิธีการค้นคำคล้ายเมื่อพบคำที่ไม่มีในพจนานุกรม

จากวิธีการหาขอบเขตของคำที่ไม่พบในพจนานุกรม ทำให้ได้คำที่ไม่พบในพจนานุกรมและทราบถึงจำนวนตัวอักษรที่ประกอบในคำที่ผิดนั้น โดยการค้นหาคำคล้ายจะทำได้ด้วยการนำจำนวนตัวอักษรที่ผิดในคำนั้นมาทำการแบ่งครึ่งทั้งทางซ้ายและทางขวา แล้วนำตัวอักษรที่ได้ทำการแบ่งมาคำนวณผ่านค่า Hashing Function ทั้งซ้ายและขวา เพื่อจะค้นหารายการคำที่น่าจะเป็นมาให้ทำการแก้ไขต่อไป เช่นจากตัวอย่างข้างต้น จะได้คำว่า "รตวจ" เป็นคำผิด ดังรูปที่ 5



หมายเหตุ ตัวอักษร " สระอำ " ในภาษาไทยจะถือว่าเป็น 1 ตัวอักษรในคำของภาษาไทย

รูปที่ 5 รูปแสดงวิธีการค้นหาคำคล้ายในพจนานุกรม

จากการค้นหาคำคล้ายจะพบว่าคำว่า "รตวจ" จะมีคำที่คล้ายกัน คือ คำว่า รตนะ, รตะ, ตรวจ, ตำรวจ, สำรวจ เป็นต้น ทำให้สามารถเลือกคำที่ถูกต้องตามพจนานุกรมในการแก้ไขคำที่ผิด

การเปรียบเทียบความสามารถของ Word กับผลงานที่ทำการวิจัย

ในประโยคภาษาไทยที่ นำมาตรวจสอบกับ Software มาตราฐานของ Word ยังพบข้อผิดพลาดในการแบ่งแยกคำอยู่ ทั้งๆที่คำในประโยคนั้นๆเป็นคำที่ถูกต้องและมีอยู่ในพจนานุกรมดังตัวอย่างนี้

ตารางที่ 5 ตารางแสดงตัวอย่างประโยคที่ Word ตัดแบ่งคำผิด

ตัวอย่างประโยค	Word 6.0	Word 7.0
การกระจาย มาบริจาค ปลากระป๋อง จะมองหาสมาชิก	การก<u>ระจาย</u> มาบ<u>ริจาค</u> ปลาก<u>ระป๋อง</u> จะมองหาส<u>มาชิก</u>	การก<u>ระจาย</u> มาบ<u>ริจาค</u> ปลาก<u>ระป๋อง</u> จะมองหาส<u>มาชิก</u>

จากตารางจะพบว่าคำที่ Word ตัดผิดนั้นเป็นคำที่เขียนถูกต้องตามพจนานุกรม เช่น ปลากระป๋อง ถ้าทำการแยกจากกันเป็น ปลา กับ กระป๋อง ใน Word จะทำถูก แต่ถ้าเขียนติดกันเป็น ปลากระป๋อง Word จะบอกว่าผิดดังตารางที่ 5 และคำที่ประกอบกันเป็นประโยคเมื่อทำการแยกเป็นคำๆจะพบว่าการตัดแบ่งคำใน Word สามารถตัดแบ่งได้ถูกต้อง นอกจากนี้ยังพบว่าคำบางคำที่ Word ไม่สามารถตัดแบ่งได้แต่คำๆนั้นเมื่อตรวจสอบในพจนานุกรมพบว่าเป็นคำที่ถูกต้อง ตัวอย่างเช่น คำว่า เศรษฐกิจ, แม่นยำ, อุตสาหกรรม เป็นต้น ในผลงานที่ได้ทำการวิจัยยังทำการเปรียบเทียบความเร็วในการตัดแบ่งคำระหว่าง Word กับ ผลงานที่ได้ทำการวิจัย ดังตารางที่ 6

ตารางที่ 6 แสดงการเปรียบเทียบความเร็วในการตัดแบ่งคำ (sec)

Software \ จำนวนหน้า / จำนวนบรรทัด	5 145	10 288	20 575	50 1436	100 2868	200 5736
Word 6.0	3	4	9	20	38	75
โปรแกรม Spell Check (พิจารณาจนสุดบรรทัด)	2	3	6	14	28	55
โปรแกรม Spell Check (พิจารณา 15 ตัวอักษร)	1	2	4	7	13	26
โปรแกรม Spell Check (พิจารณา 25 ตัวอักษร)	1	2	4	9	17	35

หมายเหตุ การทดสอบความเร็วในการตัดแบ่งใน Word 7.0 ไม่สามารถตรวจสอบได้เพราะถ้าเอกสารมีจำนวนตัวอักษรมากๆ การทำงานของ Word 7.0 จะไม่มีการตอบสนอง (Not Respond)

4. สรุปและวิจารณ์ผล

จากการใช้ Hashing Algorithm ในการตัดแบ่งแยกคำในภาษาไทย และการใช้โครงสร้างแบบ Hashing ในการเก็บคำในพจนานุกรม ทำให้ประสิทธิภาพของการตรวจสอบการตัดแบ่งคำในภาษาไทย ทำงานได้อย่างรวดเร็วกว่าการที่ใช้ Software มาตราฐาน Word มาทำการตัดแบ่งคำ มีความถูกต้อง มีความแม่นยำในการตัดแบ่งแยกคำสูง ง่ายต่อการพัฒนาและเมื่อพบคำที่ผิดในเอกสารโครงสร้างที่จัดเก็บนี้จะช่วยค้นหาคำที่คล้ายกับคำที่ผิดได้อย่าง

รวดเร็ว และยังสามารถเพิ่มคำใหม่ที่จะเกิดขึ้นต่อไปในอนาคตได้อย่างง่ายดาย

กิตติกรรมประกาศ

ผู้วิจัยและพัฒนาต้องขอขอบพระคุณอย่างสูงต่อศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ได้เอื้อเฟื้อไฟล์รายการคำของข้อมูลในพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525

เอกสารอ้างอิง

[1] James L. Peterson , " Computer Programs for Detecting and Corrceting Spelling Errors" , Communication of the ACM , Vol . 23 , No 12 , December 1980.

[2] James L. Peterson , " A Note On Undetected Typing Errors " ,Communication of the ACM , Vol 29 , No 7 , July 1986 .

[3] Jurgen Schurmann , " A Multifont Word Recognition System for Postal Address Reading " , IEEE Transaction on Computers , Vol c - 27 , No 8 , August 1978 .

[4] Rainer Hoch and Thomas Kiceinger , " On Virtual Partitioning of Large Dictionaries for Contextual Post - Processing to Improve Character Recognition", IEEE. Conference 1993 .

[5] S. Charnyapornpong, " A Thai Syallable Separation Algorithm", AIT Thesis No. CA-83-10.

[6] ยืน ภู่วรวรรณ , สมนึก คีรีโต และ สุรศักดิ์ สงวนพงษ์ "การตัดพยางค์คำไทยด้วยโครงสร้างข้อมูลแบบต้นไม้" , รายงานการประชุมวิชาการ ทางวิศวกรรมไฟฟ้า 8 สถาบันอุดมศึกษา ครั้งที่ 8 , ธันวาคม 2528.

[7] ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์ , " การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี " , รายงานการประชุมทางวิชาการ วิศวกรรมไฟฟ้า สถาบันอุดมศึกษาแห่งประเทศไทย ครั้งที่ 9 , ธันวาคม 2529 .

[8] ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์ , " การตรวจสอบตัวสะกดด้วยคอมพิวเตอร์ " , รายงานการประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 10 , พฤศจิกายน 2530.

[9] ผศ.ดร. เสรี เศวตเศรนี และ รศ. ยืน ภู่วรวรรณ , " การพัฒนาอัลกอริทึม สำหรับประมวลผลตัวอักษรไทย " , วิศวกรรมสาร ปีที่ 37 เล่มที่ 3 , มิถุนายน 2527 .

[10] " พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 ", ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC).

[11] ยืน ภู่วรวรรณ และ ชัยยงศ์ วงศ์ชัยสุวัฒน์, " การประมวลผลภาษาธรรมชาติ (Natural Language Processing) ", สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี.

[12] ฉิม ตันติยาสวัสดิกุล , " มาแบ่งคำภาษาไทยให้ดีกันเถิด " , คอมพิวเตอร์ ปีที่ 16 ฉบับที่ 84 มีนาคม - เมษายน 2532.

ประวัติผู้เขียน

ชื่อผู้เขียน	นาย สรศักดิ์ ไทยแท้
วัน เดือน ปีเกิด	30 กรกฎาคม พ.ศ. 2516
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	วท. บ. (เอกคณิตศาสตร์)
สถานที่สำเร็จการศึกษา	คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2538