บทที่ 2

แนวคิดและทฤษฎีที่เกี่ยวข้อง

การแบ่งวรรคตอนในภาษาไทยเป็นสิ่งที่หลาย ๆ คนยังมีความสับสนในบทนี้จะได้พิจารณาการแบ่งวรรคตอนที่ถูกต้องซึ่งกำหนดไว้โดยราชบัณฑิตยสถาน [4] รวมทั้งแนวทางมาตรฐานที่ควรจะเป็นของการพัฒนาโปรแกรมตัดคำภาษาไทย และ การพัฒนาการตัดคำภาษาไทยที่ผ่านมาจะได้กล่าวไว้ในบทนี้

2.1 การแยกคำหรือการตัดคำ

การแยกคำหรือการตัดคำหมายถึง การแยกคำออกเป็นคำ ๆ จากประโยคภาษาไทยที่เขียนติดกันเพื่อใช้ในการทำงานอื่น ๆ ต่อไป คำสองคำนี้สามารถใช้แทนกันได้เนื่องจากหมายถึงการทำงานอย่างเดียวกัน งานวิจัยหลาย ๆ งาน ใช้คำว่าการตัดคำ ขณะที่งานวิจัยบางงานใช้คำว่าการแยกคำ สำหรับงานวิจัยนี้จะใช้สองคำนี้ ซึ่งหมายถึงการทำงานเดียวกันสำหรับภาษาอังกฤษในหลายงานวิจัยใช้คำว่า Segmentation ขณะที่บางงานวิจัยใช้ Separation ซึ่งหมายถึงการทำงานเดียวกันเช่นกัน

2.2 เครื่องหมายและการแบ่งวรรคตอนภาษาไทยที่ถูกต้อง

ราชบัณฑิตยสถานได้กำหนด เครื่องหมาย และ การแบ่งวรรคตอนภาษาไทยที่ถูกต้อง ดังจะกล่าวอย่างสังเขปได้ต่อไปนี้คือ

2.2.1 เครื่องหมายวรรคตอนภาษาไทย ตามที่ปรากฏในเอกสารเผยแพร่ เรื่อง "หลักเกณฑ์การใช้เครื่องหมายวรรคตอน และเครื่องหมายอื่น ๆ หลักเกณฑ์การเว้นวรรคหลักเกณฑ์การเขียนคำย่อ" ของราชบัณฑิตยสถาน พ.ศ. 2530 [4] ได้รวบรวมไว้ในตาราง 2.1 และตาราง 2.2 เครื่องหมายวรรคตอนโบราณและเครื่องหมายอื่น ๆ

ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	เครื่องหมายวรรคตอน
มหัพภาค	full stop, period	.
จุลภาค	comma	,
อัฒภาค	semicolon	;
ทวิภาค	colon	:

วิภัชภาค	-	:-
ยัติภังค์	hyphen	-
นขลิขิต (วงเล็บ)	parentheses	()
วงเล็บเหลี่ยม	square brackets	[]
วงเล็บปีกกา	braces	{}
ปรัศนี	question mark	?
อัศเจรีย์	exclamation mark	!
อัญประกาศ	double quotation marks	" "
อัญประกาศเดี่ยว	single quotation marks	' '
ไม้ยมก หรือ ยมก	-	ๆ
ไปยาลน้อย หรือ เปยยาลน้อย	-	ฯ
ไปยาลใหญ่ หรือ เปยยาลใหญ่	-	ฯลฯ
ไข่ปลา หรือ จุดไข่ปลา	elipsis, dotted line	...
เส้นประ	dashed line	- - -
เสมอภาค หรือ เท่ากับ	equals	=
สัญประกาศ	underline	ขีดเส้นใต้
บุพสัญญา	ditto mark	‖
มหัตถสัญญา	-	(ย่อหน้าขึ้นบรรทัดใหม่)
ทับ	virgule, slant, slash	/

ตารางที่ 2.1 เครื่องหมายวรรคตอนภาษาไทย

ชื่อภาษาไทย	เครื่องหมายวรรคตอน
ฟองมัน หรือ ตาไก่	⓪
ฟองมันฟันหนู ฟันหนูฟองมัน หรือฝนทองฟองมัน	⊙
อังคั่นเดี่ยว ขั้นเดี่ยว หรือ ขั้น	ฯ
อังคั่นคู่ หรือ ขั้นคู่	ฯ‖
อังคั่นวิสรรชนีย์	ฯ‖:
โคมูตร (เยี่ยววัว หรือ เยี่ยวโค)	∽
ยามักการ	⸍
ทัณฑฆาต	ˊ
ตีนครุ หรือ ตีนกา	+

ตารางที่ 2.2 เครื่องหมายวรรคตอนโบราณและเครื่องหมายอื่น ๆ

2.2.2 หลักเกณฑ์การเว้นวรรคตอนภาษาไทย

2.2.2.1 กรณีที่ต้องเว้นวรรคเสมอ

2.2.2.1.1 เมื่อจบประโยคสมบูรณ์

2.2.2.1.2 ในอเนกรรถประโยค ที่มีสันธาน และ หรือ แต่ เพราะ ฯลฯ เชื่อม ให้เว้นวรรคหน้าประโยคที่ขึ้นต้นคำสันธาน

2.2.2.1.3 ระหว่างชื่อกับนามสกุล

2.2.2.1.4 ระหว่างชื่อบุคคลกับตำแหน่ง

2.2.2.1.5 ระหว่างยศกับชื่อ

2.2.2.1.6 ระหว่างตัวหนังสือกับตัวเลข

2.2.2.1.7 ระหว่างตัวหนังสือไทยกับตัวหนังสือภาษาอื่น

2.2.2.1.8 เพื่อแยกรายการต่าง ๆ

2.2.2.1.9 ระหว่างวันกับเวลา

2.2.2.1.10 ระหว่างชื่อสถานที่ต่าง ๆ

2.2.2.1.11 ระหว่างจำนวนและกลุ่มตัวเลข

2.2.2.1.12 หลังเครื่องหมายวรรคตอนและเครื่องหมายอื่น ๆ

2.2.2.1.13 ข้างหน้าและข้างหลังไปยยาลใหญ่ ยมก และ เสมอภาค

2.2.2.1.14 ข้างหลังเครื่องหมายไปยาลน้อย

2.2.2.1.15 ข้างหลังข้อความที่เป็นหัวข้อ

2.2.2.1.16 ข้างหลังหน่วยต่าง ๆ

2.2.2.1.17 ข้างหลังวลีบอกเวลาที่เป็นกลุ่มคำยาว ๆ

2.2.2.1.18 หลังคำนำหน้านามแต่ละชนิด

2.2.2.1.19 ข้างหลังคำนำหน้าพระนามพระบรมวาศานุวงศ์

2.2.2.1.20 ระหว่างพระนามกับฐานันดรศักดิ์

2.2.2.1.21 ระหว่างชื่อบริษัท ธนาคาร ฯลฯ กับคำ “จำกัด” ที่อยู่ท้ายชื่อ

2.2.2.1.22 ระหว่างคำ “ห้างหุ้นส่วนจำกัด” และ “ห้างหุ้นส่วนนิติบุคคล” กับชื่อ

2.2.2.1.23 ข้างหน้า และข้างหลังคำ ณ ธ

2.2.2.1.24 ข้างหน้า และข้างหลังคำ “ได้แก่”

2.2.2.1.25 หน้าคำสันธาน และ หรือ ในรายการ

2.2.2.1.26 ข้างหน้าคำ “เป็นต้น”

2.2.2.1.27 ข้างหลังคำ “ว่า”

2.2.2.2 กรณีที่ไม่เว้นวรรค

2.2.2.2.1 ไม่เว้นวรรคระหว่าง คำนำหน้าชื่อ กับชื่อ

2.2.2.2.2 ไม่เว้นวรรคระหว่าง บรรดาศักดิ์ สมณศักดิ์ ฐานันดรศักดิ์ กับ นาม หรือราชทินนาม

2.2.2.2.3 ไม่เว้นวรรคระหว่างคำนำหน้าชื่อ ที่เป็นตำแหน่งหรืออาชีพ กับชื่อ

2.2.2.2.4 ไม่เว้นวรรคระหว่างคำนำหน้าชื่อที่แสดงฐานะของนิติบุคคล หน่วยงาน หรือกลุ่มบุคคลกับชื่อ

2.3 การพัฒนาและขั้นตอนวิธีการตัดคำภาษาไทยแบบต่าง ๆ

ความคืบหน้าของงานวิจัยสาขานี้ยังไม่ถึงจุดสุดท้าย หน่วยงานวิจัยของภาครัฐและเอกชน มีการดำเนินการวิจัยอย่างต่อเนื่องเพื่อแก้ปัญหาการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ การวิเคราะห์ประเมินปัญหาที่สมบูรณ์ยังไม่ปรากฏชัดเจน การเปรียบเทียบประสิทธิภาพของการตัดคำแบบต่าง ๆ จะเป็นเครื่องมือวัดที่จะทำให้เห็นปัญหาที่ซ่อนเร้นอยู่ อีกทั้งเห็นแนวทางที่ถูกต้องที่จะทำการวิจัยในสาขานี้ต่อ

2.3.1 ลักษณะพื้นฐานของประโยคภาษาไทย

ประโยคในภาษาไทย จะประกอบไปด้วยคำ หรือ กลุ่มของคำหลาย ๆ คำก็ได้ จะได้อธิบายในรายละเอียดดังต่อไปนี้คือ เมื่อให้

ST แทนประโยค
Si แทนกลุ่มของคำ
i แทนลำดับที่ของกลุ่มคำ
S1, S2, S3, ..., Sm คือกลุ่มของคำที่เป็นลำดับที่ 1, 2, 3, ..., m ในประโยค
Wij แทนคำศัพท์ที่เป็นสมาชิกในกลุ่มคำศัพท์ Si ลำดับที่ j

จะได้ว่า

$$ST = [S1, S2, S3, ..., Sm-1, Sm]$$

เมื่อ

Si = [Wi1, Wi2,..., Win]

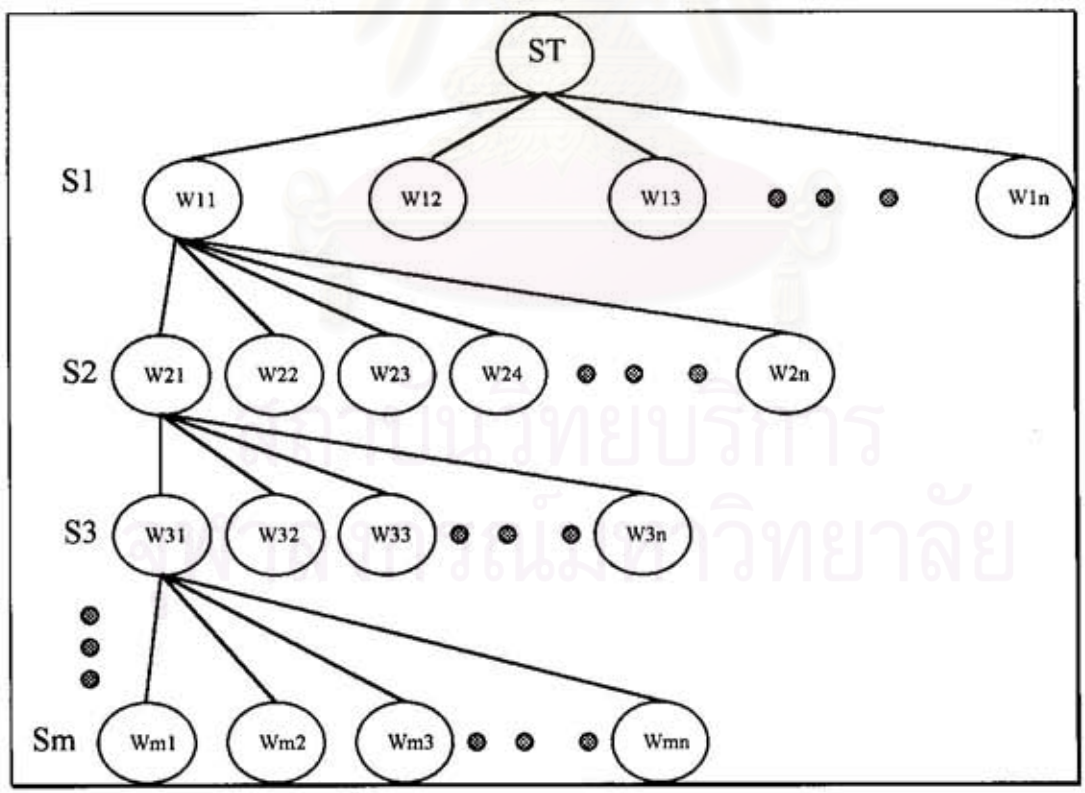
ถ้าเขียนเป็นแผนภูมิจะได้ดังรูป 2.1 ซึ่งถ้าทำการตัดคำออกมาจากประโยคภาษาไทยแล้วจะเป็นไปได้ว่าเกิดมีประโยคขึ้นมามากกว่าหนึ่งประโยค ตัวอย่างเช่น พิจารณาประโยคข้างล่างนี้

ฉันมารอกราบ

เมื่อผ่านการตัดคำแล้วอาจจะไได้ประโยคดังต่อไปนี้

ฉัน-มา-รอ-กราบ
ฉัน-มา-รอก-ราบ
ฉัน-มาร-อก-ราบ

สิ่งนี้เป็นปัญหาหนึ่งที่ขั้นตอนวิธีตัดคำแต่ละแบบตัดคำแล้วได้ระดับความถูกต้องของประโยค (syntax validity) ที่แตกต่างกันออกไป



รูป 2.1 ลักษณะพื้นฐานของประโยคภาษาไทย

2.3.2 วิธีการใช้กฎเกณฑ์ (rules based)

ขั้นตอนวิธีการตัดคำ ที่คิดค้นพัฒนาในยุกแรก ๆ จะใช้วิธีการตรวจสอบกฎเกณฑ์ต่าง ๆของคำภาษาไทยที่ประกอบขึ้นเป็นประโยค เช่นกฎเกณฑ์ของตัวอักษรที่อยู่ติดกัน หรือกฎเกณฑ์ที่กำหนดโดยราชบัณฑิตยสถาน ยกตัวอย่างเช่น

2.3.2.1 สระหน้า (เ แ โ ไ ใ) เป็นตำแหน่งเริ่มของคำเสมอ

2.3.2.2 ทันฑฆาต (์) เป็นตำแหน่งจบคำเสมอ

2.3.2.3 ไม่ตัดคำหน้าวรรณยุกต์และสระตาม (ะ ั า ำ ๆ ิ ี ึ ื ุ ู ็)

2.3.2.4 อักขระหรือเครื่องหมายแบ่งวรรคตอนเป็นตำแหน่งแบ่งคำเสมอ

2.3.2.5 สามารถตัดคำได้ระหว่างตัวหนังสือกับตัวเลขเสมอ

2.3.2.6 สามารถตัดคำได้ ระหว่างตัวหนังสือไทย กับตัวหนังสือต่างประเทศเสมอ เป็นต้น

ซึ่งวิธีการแบบนี้จะมีข้อจำกัดมากนั่นคือผลของการตัดคำอาจได้เป็นกลุ่มของคำ ซึ่งในความเป็นจริงยังสามารถตัดคำแยกย่อยออกไปได้อีกแต่จะมีประสิทธิภาพในเชิงความเร็วที่สูง การใช้งานทรัพยากรน้อย ความถูกต้องของประโยคค่อนข้างดี แต่ความถูกต้องของคำหลังการตัดต่ำ ซึ่งสามารถใช้ได้ในงานบางประเภทที่ไม่ต้องการตัดแยกคำออกให้ย่อยที่สุด เช่นงานจัดรูปแบบของเอกสารแต่จะไม่เหมาะกับงานที่ต้องการหน่วยคำที่แยกย่อยที่สุดเช่นงานตรวจสอบตัวสะกด หรืองานแปลภาษาด้วยเครื่องเป็นต้น

2.3.3 วิธีการใช้พจนานุกรมช่วย (dictionary approach)

ในยุคต่อ ๆ มาขั้นตอนวิธีการตัดคำภาษาไทยโดยส่วนใหญ่ จะใช้พจนานุกรมเข้าช่วย และอาจจะมีเทคนิคที่เพิ่มเติมเข้ามา หรืออาจผสมผสานกับวิธีการใช้กฎเกณฑ์ งานวิจัยของสมปราถนา รัทยานนท์ [1] ได้เสนอการนำพจนานุกรมช่วยในการตัดคำภาษาไทย โดยค้นหาศัพท์จากประโยคที่ต้องการตัดคำ แล้วนำคำศัพท์ที่ค้นหาได้จัดเก็บในอะเรย์ชุดหนึ่ง จากนั้นจึงสร้างอะเรย์ของคำศัพท์ชุดต่อไป ด้วยการค้นหาศัพท์จากประโยคอินพุทที่ได้ตัดคำศัพท์ที่เจอก่อนหน้าไปแล้ว ซึ่งจะกระทำกับคำศัพท์ทุกคำที่พบเจอในประโยค ผลที่ได้อาจจะเกิดประโยคมากกว่าหนึ่งประโยคก็ได้ ขั้นตอนการตัดคำสามารถอธิบายได้ดังต่อไปนี้

2.3.3.1 สร้างอะเรย์ของคำศัพท์ ที่ค้นหาจากพจนานุกรม ของ ประโยคเดิมจากคำแรกที่พบเจอทั้งหมด

2.3.3.2 ดึงคำศัพท์ในอะเรย์ชุดแรกมา แล้ว สร้างอะเรย์ของคำศัพท์ที่ค้นหาได้จากพจนานุกรม ของประโยคอินพุทประโยคใหม่ ที่ตัดคำศัพท์ที่เจอครั้งแรก ออกไป จนกระทั่งสิ้นสุดประโยคและให้อะเรย์ของคำชุดสุดท้ายที่สร้างขึ้นเป็นอะเรย์ที่ n

2.3.3.3 ย้อนการทำงานแบบเดิมกับอะเรย์ชุดที่ n-1 จนถึง ชุดที่ 1 นั่นคือทำกับคำศัพท์ทุกคำที่พบเจอในประโยค ซึ่งบางครั้งในหนึ่งประโยคจะตัดคำแล้วได้หลายประโยคได้

วิธีการนี้จะสิ้นเปลืองทรัพยากรหน่วยความจำหลักค่อนข้างมาก เนื่องจากทั้งพจนานุกรมและอะเรย์ของคำศัพท์ที่ตัดแล้วจะสร้างและเก็บไว้ในนั้นทั้งหมด ทั้งประสิทธิภาพเชิงความเร็วและความถูกต้องของคำจะขึ้นอยู่กับเทคนิคที่ใช้ในการค้นหาคำจากพจนานุกรม และปริมาณคำศัพท์ที่มีอยู่ในพจนานุกรมตามลำดับ และปริมาณของกลุ่มคำในประโยคจะแปรผันตรงกับ ประสิทธิภาพเชิงความเร็วด้วย ไม่สามารถเลือกประโยคที่มีความถูกต้อง ทางไวยากรณ์ จากหลาย ๆ ประโยคที่ตัดมาได้ แต่ถ้ามีปริมาณของคำศัพท์ในพจนานุกรมที่เหมาะสม วิธีการนี้จะได้ความถูกต้องของคำหลังการตัด (word validity) สูงวิธีหนึ่งทีเดียว

การออกแบบพจนานุกรมจะมีผลมากต่อประสิทธิภาพเชิงความเร็ว และการใช้งานทรัพยากรของขั้นตอนวิธีการตัดคำแบบนี้ งานวิจัยของ ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์ [3] ได้เสนอโครงสร้างข้อมูลของพจนานุกรมไว้สามแบบ ที่ทำให้การค้นหาทำได้รวดเร็วและประหยัดเนื้อที่ หรือลดความซ้ำซ้อนของข้อมูลที่เก็บอยู่ในพจนานุกรมคือ

ก. แบบแรก Table-Table-Linear-Search จะใช้สองตัวอักษรแรกของคำมาทำตารางดรรชนีสองระดับที่ชี้ไปยังคำที่จัดเรียงติดกันไปในพจนานุกรม โดยมีความยาวของแต่ละคำกำกับ จำนวน 5400 คำ โดยที่จะไม่เก็บคำที่ประกอบขึ้นมาจากหน่วยคำย่อย เช่นคำว่า “ไฟฟ้า” ประกอบขึ้นมาจากคำว่า “ไฟ” กับ “ฟ้า” เป็นต้นเพื่อความประหยัดเนื้อที่เก็บ

ข. แบบที่สอง Table-Index-Search ใช้ตัวอักษรแรกเป็นดรรชนีระดับที่หนึ่งบอกขอบเขตของกลุ่มอักษร ดรรชนีระดับที่สองเก็บตำแหน่งของคำในพจนานุกรมที่เก็บคำทั้งหมดเรียงกันเป็นสตริง

ค. แบบที่สาม Tree Structure แต่ละโหนดจะเก็บตัวอักษรโดยมีกิ่งก้านเชื่อมโยงเป็นคำ และสิ้นสุดแต่ละคำที่ไป

งานวิจัยดังกล่าวได้ทำการทดลอง และวัดผลการใช้งานพจนานุกรมทั้งสามแบบเปรียบเทียบกัน ซึ่งสรุปผลได้ว่าแบบที่สองจะมีประสิทธิภาพเชิงความเร็วของการเปรียบเทียบคำสูงสุด แต่ประสิทธิภาพการใช้งานทรัพยากรด้อยที่สุด ในขณะที่แบบแรกจะมีประสิทธิภาพของการใช้ทรัพยากรที่ดีที่สุด

2.3.4 วิธีการเทียบคำที่ยาวที่สุด (longest word pattern matching)

งานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์ และ ดร.จงกล งามวิวิทย์ และคณะ [5] ได้นำเสนอขั้นตอนวิธีการตัดคำวิธีนี้ไว้ ซึ่งเป็นวิธีการหนึ่งที่ต้องใช้พจนานุกรมช่วยเช่นกัน เริ่มต้นจากการใช้รหัสแอสกีของตัวอักษรแรกของหน่วยคำที่จะแยกจากประโยค เป็นดัชนีในการเลือก

กลุ่มของดัชนีคำศัพท์ในพจนานุกรม จากนั้นก็หาค่าน้ำหนัก (weight) ของหน่วยคำนั้น เพื่อเป็นดัชนีตัวต่อไปในการเลือกกลุ่มย่อยในดัชนีคำศัพท์นั้น ๆ พร้อมกับเลือกกลุ่มคำศัพท์ที่มีความยาวเท่ากับความยาวของหน่วยคำ ที่จะแยกออกจากประโยค เพื่อทำการเปรียบเทียบต่อไป โดยที่การจัดเรียงโครงสร้างคำศัพท์ในพจนานุกรม สำหรับขั้นตอนวิธีการตัดคำแบบนี้สามารถเขียนเป็นสมการคณิตศาสตร์ได้ดังนี้

$$B = [Cj\ [Dj,\ k[Ej,\ m[F]]]]$$

เมื่อ

B เป็นกลุ่มของคำศัพท์ที่มีอยู่ในภาษาไทยทั้งหมด และมีสมาชิก Cj แล้ว

Cj หมายถึงกลุ่มค่ารหัสแอสกีของอักขระตัวแรกของคำศัพท์ ที่มีอยู่ในภาษาไทยซึ่งจะเรียกว่ากลุ่มดัชนีของคำศัพท์ และ j หมายถึงลำดับของกลุ่มดัชนีคำศัพท์ที่มีอยู่ 44 กลุ่ม แบ่งเป็นพยัญชนะ 39 กลุ่มได้แก่ดังปรากฏในตาราง 2.3 และกลุ่มของสระหน้าอีก 5 กลุ่มดังในตาราง 2.4

C1 = 161(ก)	C2 = 162(ข)	C3 = 164(ค)	C4 = 166(ฆ)	C5 = 167(ง)
C6 = 168(จ)	C7 = 169(ฉ)	C8 = 170(ช)	C10 = 172(ฌ)	C10 = 172(ฌ)
C11 = 173(ญ)	C12 = 174(ฎ)	C13 = 175(ฏ)	C14 = 176(ฐ)	C15 = 179(ณ)
C16 = 180(ด)	C17 = 181(ต)	C18 = 182(ถ)	C19 = 183(ท)	C20 = 184(ธ)
C21 = 185(น)	C22 = 186(บ)	C23 = 187(ป)	C24 = 188(ผ)	C25 = 189(ฝ)
C26 = 190(พ)	C27 = 191(ฟ)	C28 = 192(ภ)	C29 = 193(ม)	C30 = 194(ย)
C31 = 195(ร)	C32 = 196(ฤ)	C33 = 197(ล)	C34 = 199(ว)	C35 = 200(ศ)
C36 = 202(ส)	C37 = 203(ห)	C38 = 205(อ)	C39 = 206(ฮ)	

ตาราง 2.3 กลุ่มค่ารหัสแอสกีของอักขระตัวแรก

C40 = 224(เ)	C41 = 225(แ)	C42 = 226(โ)	C43 = 227(ใ)	C44 = 228(ไ)

ตาราง 2.4 กลุ่มค่ารหัสแอสกีของสระหน้า

ซึ่งในแต่ละกลุ่มของดัชนีคำศัพท์จะมี Dj, k เป็นสมาชิกซึ่ง Dj, k หมายถึงกลุ่มของผลรวมระหว่างค่ารหัสแอสกี ของอักขระตัวแรกกับ รหัสแอสกี ของอักขระตัวที่สอง ซึ่งเรียกว่า

น้ำหนัก (Weight) โดยมี k เป็นลำดับของกลุ่มนี้มีสมาชิกเป็น Ej, m ซึ่งคือกลุ่มของคำศัพท์ (F) ที่มีจำนวนอักขระเท่ากัน

การเปรียบเทียบคำศัพท์ที่เก็บในพจนานุกรมกับประโยคที่ต้องการจะแยกแยะคำโดยวิธีการเปรียบเทียบคำที่ยาวที่สุดนี้ จะทำตามขั้นตอนดังต่อไปนี้

2.3.4.1 หาค่ารหัสแอสกีของอักขระตัวแรก น้ำหนักของประโยค และความยาวของประโยคที่ต้องการจะตัดคำ

2.3.4.2 เข้าหากลุ่มของดัชนีคำศัพท์ ที่มีรหัสแอสกีเท่ากับรหัสแอสกีของอักขระแรกของประโยคนั้น

2.3.4.3 เข้าหากลุ่มย่อยต่อไป โดยเปรียบเทียบกับน้ำหนักของดัชนีย่อยของคำศัพท์

2.3.4.4 นำความยาวของประโยคมาเปรียบเทียบกับค่า Ej, m

2.3.4.5 ถ้าความยาวของประโยค มากกว่า ความยาวของคำศัพท์ ที่ยาวที่สุดของกลุ่มคำศัพท์นั้น ให้เปรียบเทียบจากคำศัพท์ที่ยาวที่สุดไปหาคำศัพท์ที่สั้นที่สุด

2.3.4.6 ถ้าความยาวของประโยคเท่ากับความยาวของคำศัพท์กลุ่มใด ๆ ให้เริ่มเปรียบเทียบกลุ่มย่อยไปหากลุ่มคำศัพท์ที่สั้นสุด

2.3.4.7 ถ้าความยาวของประโยค น้อยกว่าความยาวของคำศัพท์ ก็ให้เริ่มเปรียบเทียบกับกลุ่มคำศัพท์ถัดไปที่มีความยาวน้อยกว่า ไปจนถึงกลุ่มคำศัพท์ที่สั้นที่สุด

2.3.4.8 เปรียบเทียบคำศัพท์ อักขระ ต่ออักขระ ถ้าไม่มีผลต่าง ก็ทำเครื่องหมายสามารถตัดคำออกจากประโยคได้ และนำคำศัพท์อื่น ๆ ที่อยู่ในกลุ่มนั้นมาเปรียบเทียบต่อจนถึงคำศัพท์สุดท้าย ที่มีความยาวน้อยที่สุด แล้วตัดคำออกจากประโยค ตามจุดที่ทำเครื่องหมายไว้

2.3.4.9 ไม่ว่าจะตัดได้คำเดียว หรือหลายคำ ให้นำประโยคที่เหลือจากการตัดมาทำการเปรียบเทียบต่อไปจนจบประโยค

2.3.4.10 ถ้าไม่พบเจอคำศัพท์เลยในกลุ่มดัชนีย่อย ให้ยกเลิกการทำงาน และถือว่าประโยคที่นำมาตัดคำนั้น ไม่ถูกต้อง

2.3.5 วิธีการย้อนกลับ (back tracking) เป็นเทคนิคที่นำมาช่วยในการตัดคำในกรณีที่เกิดคำที่ไม่ถูกต้องขึ้นมา หลังจากผ่านขั้นตอนวิธีการตัดคำนั้น ๆ โดยทำการย้อนกลับเรื่อย ๆจนสามารถแก้ไขคำจนถูกต้อง หรือไม่สามารถทำต่อไปได้อีกต่อไป ซึ่งอาจจะเป็นไปได้ว่ามีคำที่เขียนผิดปรากฏอยู่

งานวิจัยของบุญเรือง ธนาสุนทรไพศาล [2] ได้ให้แนวทางการใช้การตัดคำร่วมกันในระหว่างโปรแกรมประยุกต์หลาย ๆ โปรแกรม ว่าจะต้องมีส่วนเชื่อมต่อระหว่างส่วนโปรแกรมการตัดคำกับโปรแกรมประยุกต์ที่ต้องการใช้การตัดคำนั้นโดยที่ส่วนการเชื่อมต่อจะมีข้อมูลที่เป็นกลางที่สุดสำหรับทุก ๆโปรแกรมประยุกต์ และต้องรองรับได้ทั้งการตัดคำ แบบโต้ตอบ หรือ แบบ

ข้อความ เป็นกลุ่ม โดยที่ส่วนเชื่อมต่อจะส่งค่ากลับเป็นตำแหน่งของการตัดคำหรือบอกว่ายังไม่สามารถตัดได้กับข้อมูลที่มีอยู่นี้

งานวิจัยนี้เป็นตัวอย่างของความพยายามให้เกิดความเป็นมาตรฐานและการที่จะใช้ขั้นตอนการตัดคำร่วมกันในโปรแกรมประยุกต์หลาย ๆโปรแกรม ซึ่งผู้พัฒนาส่วนจัดการภาษาไทยน่าจะได้กำหนดให้ ขั้นตอนวิธีการตัดคำ เป็นการให้บริการมาตรฐานพื้นฐานที่มีอยู่ สำหรับโปรแกรมประยุกต์ทั่ว ๆไปโดยประกาศมาตรฐานการเชื่อมโยงขึ้นมา

2.4 การวิเคราะห์เลือกประโยคที่ถูกต้องหลังการตัดคำ

เนื่องจากผลของการตัดคำแบบต่าง ๆ จะทำให้ได้ประโยคมากกว่าหนึ่งประโยค จึงเกิดความจำเป็น ในการเลือกประโยคที่มีความถูกต้องที่สุด ในแง่ของไวยากรณ์ และในแง่ของความหมายของประโยคที่ได้

มีงานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์ และคณะ [6] ได้นำเสนอวิธีการแก้ปัญหานี้โดยอาศัยข้อมูลทางด้านความถี่ของการใช้คำไทย (word usage frequency) เพื่อคิดเป็นความน่าจะถูกใช้ในภาษาไทย (probability of usage) ใช้สัญลักษณ์ Pu ซึ่งจะต้องพิจารณาจากตัวอย่างการใช้ประโยคจากแหล่งข่าวสารต่าง ๆ ซึ่งทำการสุ่มตัวอย่างมามาก ๆ บรรจุเข้าไปในคอมพิวเตอร์ แล้วทำการตรวจสอบจำนวนครั้งของการปรากฏของคำต่าง ๆ ความน่าจะเป็นของการใช้คำสามารถเขียนเป็นสมการได้ดังนี้

$$Pu(W1) = f1/N$$

เมื่อ Pu(W1) เป็นค่าความน่าจะถูกใช้ในภาษาประจำวันของคำไทย W1

f1 เป็นความถี่หรือจำนวนครั้งของคำไทย W1 ที่ปรากฏในคลังข้อมูลของประโยค

N เป็นจำนวนคำทั้งหมดที่บรรจุอยู่ในคลังข้อมูลของประโยค

การเลือกประโยคที่ถูกต้องจะใช้วิธีการเปรียบเทียบค่า Pu ของคำที่ลำดับเดียวกันในประโยคต่าง ๆที่ตัดคำมาได้ถ้ามีค่าที่มากกว่าก็ให้เลือกประโยคนั้น ถ้ามีค่าเท่ากันให้เปรียบเทียบคำในคู่ลำดับต่อไป

ขั้นตอนวิธีการตัดคำจากประโยคภาษาไทย โดยส่วนมากจะได้ประโยคมากกว่าหนึ่งประโยคได้สำหรับประโยคใด ๆ เนื่องจากคุณลักษณะพื้นฐานของภาษาไทย ดั้งนั้นจึงมีความจำเป็นที่จะต้องมีแนวทางในการที่จะวิเคราะห์เลือกประโยคที่ถูกต้องหลังการตัดคำมาร่วมด้วยจึงจะได้ผลการทำงานที่สมบูรณ์ และสามารถนำผลการทำงานไปใช้กับการประมวลผลอื่น ๆได้

ในบทนี้ได้ทำการรวบรวมแนวคิดและการวิจัยที่ผ่านมาในแขนงนี้ สำหรับบทต่อไปจะได้พูดถึงเครื่องมือที่จะนำมาใช้ในการวัดเปรียบเทียบประสิทธิภาพของโปรแกรมตัดคำภาษาไทยแบบต่าง ๆ