บทที่ 1

บทนำ

ความเป็นมาของปัญหา

ในปัจจุบันการจัดพิมพ์งานต่างๆที่มีสัญลักษณ์ทางคณิตศาสตร์เข้ามาเกี่ยวข้องด้วยนั้นคุณภาพของงานที่ออกมาจะไม่สวยงาม บางครั้งอาจทำให้ผู้พบเห็นมีความเข้าใจคลาดเคลื่อนจากความต้องการของผู้จัดพิมพ์ถึงแม้ว่าจะมีโปรแกรมประมวลผลคำต่างๆให้เลือกใช้มากมายแต่ก็ยังไม่มีประสิทธิภาพดีเท่าไรนัก นอกจากนี้โปรแกรมประมวลผลคำต่างๆดังกล่าวนี้มักจะถูกจำกัดด้วยกฎหมายทางลิขสิทธิ์ ทำให้ต้องเสียค่าใช้จ่ายสูงขึ้น

ในปี ค.ศ.1977 ศาสตราจารย์ โดนัลด์ อี.เคนูธ (Donald E. Knuth) ได้ตระหนักถึงปัญหาการพิมพ์หนังสือที่เกี่ยวข้องกับผลงานทางวิทยาศาสตร์ด้วยโปรแกรมประมวลผลคำดังกล่าวจะมีคุณภาพต่ำไม่เป็นที่น่าพอใจ ดังนั้นเขาจึงออกแบบตัวอักษรใหม่โดยใช้โปรแกรมเมตาฟอนต์ (METAFONT) หลังจากนั้นก็ได้พัฒนาโปรแกรมเท็กซ์ (TeX) ขึ้นเพื่อใช้กับตัวอักษรที่สร้างขึ้นด้วยโปรแกรมเมตาฟอนต์ดังกล่าว ซึ่งโปรแกรมเท็กซ์ที่พัฒนาขึ้นนี้จะให้ตัวอักษรที่มีคุณภาพดีกว่ามาก หลังจากที่การพัฒนาโปรแกรมเท็กซ์ได้สิ้นสุดลง เขายังได้อุทิศโปรแกรมนี้ให้สามารถใช้ได้ทั่วไปโดยไม่คิดค่าลิขสิทธิ์แต่อย่างใด ด้วยเหตุผลดังกล่าวนี้ ทำให้เท็กซ์มีการพัฒนาต่อไปเพื่อให้สามารถใช้งานได้กับหลายๆภาษาเช่น เยอรมัน ญี่ปุ่น เป็นต้น

สำหรับในประเทศไทยของเรา การใช้งานเท็กซ์ยังไม่แพร่หลายนัก คงใช้กันแต่ในกลุ่มของนักวิจัยเท่านั้นทั้งนี้เนื่องจากการใช้งานโปรแกรมเท็กซ์ค่อนข้างยุ่งยากและซับซ้อนกว่าการใช้โปรแกรมประมวลผลคำทั่วไปอีกทั้งยังไม่มีการพัฒนาให้สามารถแสดงผลเป็นภาษาไทยได้

ในปัจจุบันกลุ่มนักเรียนไทยที่ศึกษาอยู่ที่สถาบันเทคโนโลยีแห่งโตเกียว (Tokyo Institute of Technology) ประเทศญี่ปุ่นได้ร่วมกันศึกษาและพัฒนาโปรแกรมเท็กซ์ภาษาไทยขึ้นโดยอาศัยชุดคำสั่งต่างๆที่มีมากับตัวโปรแกรมลาเท็กซ์ (LaTeX) นั้นทำการจัดบรรทัดเพื่อแสดงผลเป็นภาษาไทย และสร้างโปรแกรมที่

สามารถอ่านเอกสารที่เป็นภาษาไทยที่สร้างขึ้นโดยโปรแกรมบรรณาธิกรภาษาไทยต่างๆ แปลงเอกสารภาษาไทยเหล่านั้นให้อยู่ในรูปแบบของแฟ้มที่มีนามสกุลเป็น .tex แล้วสามารถใช้กับโปรแกรมเท็กซ์ธรรมดาได้

เนื่องจากการพัฒนาดังกล่าวยังไม่สามารถที่จะใช้งานได้อย่างสมบูรณ์แบบ จึงยังไม่สามารถนำมาใช้งานได้อย่างจริงจัง ยังมีข้อบกพร่องที่ต้องปรับปรุงอีกมากเช่นในเรื่องของการตัดคำภาษาไทยที่ยังไม่มีการใช้โปรแกรมตัดคำที่มีประสิทธิภาพดีเท่าที่ควร ในเรื่องของตัวอักษรภาษาไทยที่ยังไม่มีการพัฒนาขึ้นมาใช้เองด้วยโปรแกรมเมตาฟอนต์แต่ต้องใช้ฟอนต์ที่พัฒนาขึ้นจากระบบอื่นแทน เหล่านี้เป็นต้น การศึกษาและแก้ไขปัญหาเหล่านี้จะทำให้โปรแกรมเท็กซ์สำหรับภาษาไทยสมบูรณ์ยิ่งขึ้น

ด้วยเหตุผลดังกล่าวข้างต้นนี้ การสร้างโปรแกรมแสดงผลตัวอักษรไทยบนเท็กซ์จะเป็นประโยชน์สำหรับงานพิมพ์ทางด้านวิชาการที่จำเป็นต้องใช้ภาษาไทยซึ่งจะแพร่หลายต่อไปในอนาคต

อนึ่งเพื่อความสะดวกในการอ้างอิง จะใช้ชื่อเรียกโปรแกรมแสดงผลตัวอักษรภาษาไทยบนระบบเท็กซ์นี้ว่า ไทยเท็กซ์ (ThaiTeX)

<u>แนวคิด ทฤษฎี หรือสมมุติฐาน ประกอบการวิจัย</u>

การทำงานของโปรแกรมเท็กซ์นั้นโดยทั่วไปแล้วมีหลักการคือจะสร้างเอกสาร(document)ขึ้นมาจากโปรแกรมบรรณาธิกรใดก็ได้ โดยเอกสารที่สร้างขึ้นมานั้นจะต้องมีรูปแบบ (syntax) ตามกฎเกณฑ์ของเท็กซ์ หลังจากนั้นก็จะใช้โปรแกรมเท็กซ์จัดการแปลงเอกสารนั้นให้เป็นแฟ้มข้อมูลที่สามารถแสดงผลตามที่ได้ออกแบบไว้

ความหมายของเอกสาร (document) ก็คือข้อมูลที่ต้องการพิมพ์ผสมกับคำสั่งต่างๆที่มีในเท็กซ์ ซึ่งหลักเกณฑ์นี้จะต่างกับโปรแกรมประมวลผลคำประเภท WYSIWYG (What You See Is What You Get) ที่ใช้กันอยู่ทั่วไปซึ่งจะมีเฉพาะเนื้อข้อมูลที่เราต้องการพิมพ์ล้วนๆเท่านั้น หน้าที่ของเท็กซ์คือการแยกระหว่างเนื้อข้อมูลที่ต้องการจะพิมพ์กับคำสั่งที่ควบคุมการทำงาน ดังนั้นถ้าต้องการตัวอักษร "A" ที่เป็นตัวหนา กับตัวอักษร "A" ที่เป็นตัวเอียง สิ่งที่ต้องทำคือพิมพ์ข้อมูลเป็นตัว "A" เช่นเดิม แต่จะต่างกันตรงคำสั่งควบคุมการทำงานของเท็กซ์ดังนี้

\bf A จะได้ตัวอักษรที่เป็นตัวหนา **A**

\sl A จะได้ตัวอักษรที่เป็นตัวเอียง *A*

นอกจากนี้เท็กซ์ยังสามารถรับข้อมูลที่กำหนดให้อยู่ในรูปแบบของรหัสแอสกีได้อีกด้วยเช่นแทนที่จะพิมพ์ A ก็อาจจะพิมพ์ 41 ผลลัพธ์ที่ได้ก็ยังคงเป็นเช่นเดิม

ด้วยหลักการดังกล่าวนี้การสร้างโปรแกรมไทยเท็กซ์จึงมีความเป็นได้โดยที่สร้างตัวอักษรไทยขึ้นและแทนตัวอักษรที่สร้างขึ้นนั้นด้วยรหัสเช่นเดียวกับการแทนตัวอักษรในภาษาอังกฤษด้วยรหัสแอสกี โดยรหัสของอักขระภาษาไทยที่ใช้นี้จะเป็นเช่นเดียวกับตัวรหัสอักขระภาษาไทยที่ใช้กันอยู่โดยทั่วไป เมื่อนำรหัสของอักขระเหล่านี้มาจับคู่เข้ากับรูปภาพของตัวอักษรที่สร้างขึ้นโดยโปรแกรมเมตาฟอนต์ โดยทำการแทรกคำสั่งของเท็กซ์เพื่อจัดการกับรหัสของตัวอักษรภาษาไทยเหล่านั้นก็จะสำเร็จเป็นโปรแกรมไทยเท็กซ์

การพัฒนาโปรแกรมไทยเท็กซ์จะสามารถแบ่งเป็นขั้นตอนการทำงานต่างๆได้ดังนี้

1. สร้างตัวอักษรภาษาไทยสำหรับใช้ในการแสดงผล โดยใช้โปรแกรมเมตาฟอนต์สร้างแฟ้มที่มีนามสกุล .mf ตามที่ทำการออกแบบไว้ให้เป็นแฟ้มข้อมูลในรูปแบบที่โปรแกรมไดรว์เวอร์ (driver program) ของอุปกรณ์การแสดงผลแต่ละชนิดสามารถนำไปเปลี่ยนเป็นพิกเซลสำหรับแต่ละ อุปกรณ์ได้

เนื่องจากตัวอักษรแต่ละตัวก็คือรูปภาพที่ประกอบไปด้วยเส้นตรง เส้นโค้ง วงกลม วงรีต่างๆ ดังนั้นถ้าสามารถกำหนดจุดต่างๆให้เหมาะสมเพื่อสร้างส่วนประกอบย่อยๆเหล่านี้ เมื่อนำองค์ประกอบย่อยมาประกอบเข้าด้วยกันก็จะได้ตัวอักษรตามต้องการ

การสร้างองค์ประกอบของตัวอักษรดังกล่าวนี้จะใช้วิธีการแทนที่ด้วยเวคเตอร์ การนำองค์ประกอบมารวมกันก็คือการนำเอาเวคเตอร์เหล่านี้มาวางไว้ที่ตำแหน่งต่างๆ อย่างเหมาะสม ดังนั้นตัวอักษรตัวเดียวกันไม่ว่าจะมีขนาดใดก็ตามจะประกอบด้วยเวคเตอร์กลุ่มเดียวกันเสมอแต่ เวคเตอร์เหล่านั้นจะมีขนาดไม่เท่ากันการเปลี่ยนขนาดของตัวอักษรก็ทำได้โดยเปลี่ยนขนาดของเวคเตอร์เท่านั้น วิธีการทำเช่นนี้จะทำให้ตัวอักษรที่ได้มีคุณภาพเหมือนเดิมเสมอไม่ว่าจะขยายหรือ ย่อขนาดลงเท่าใดก็ตาม

ในหลักการทำงานของโปรแกรมเท็กซ์นั้นจะใช้หลักการของกล่องและกาว (box and glue) ซึ่งก็คือการนำเอาตัวอักษรแต่ละตัวมาเรียงต่อกัน โดยไม่จำกัดว่าจะต้องเรียงให้อยู่ในแนวเดียวกันหรือไม่ ตัวอักษรแต่ละตัวก็คือกล่อง 1 กล่อง หรืออาจจะนำกล่องเล็กๆเหล่านี้มาเรียงต่อๆกันแล้วทำการกำหนดเป็นกล่องที่ใหญ่ขึ้นก็ได้ การเรียงต่อกันก็คือการนำมาติดกาวที่ด้านหลังและแปะลงไปบนกระดาษเปล่านั่นเอง แต่กาวนี้จะมีความยืดหยุ่นของตัวเองโดยสามารถที่จะหดแคบลง หรือขยายขนาดให้กว้างขึ้นได้ตามความต้องการของผู้ออกแบบการที่จะนำสระหรือวรรณยุกต์ไปเรียงกันเหนือหรือใต้แนวบรรทัดปกติก็จำเป็นต้องใช้หลักการนี้ กล่าวคือเท็กซ์จะไม่สนใจว่าตัวอักษรในกล่องนั้นจะมีหน้าตาเป็นอย่างไร เพียงแต่จะทำหน้าที่วางกล่องไว้ในตำแหน่งที่เหมาะสมเท่านั้น ดังนั้นการออกแบบตัวอักษรในกล่องจึงสามารถทำได้ตามความต้องการ

การกำหนดกล่องนั้นจะมี3 มิติคือ ความกว้าง (width) ความสูง (height) และ ความลึก (depth) ความกว้างคือระยะทางตามแนวระดับ (horizontal dimension) ความสูง คือระยะทางตามแนวดิ่งเหนือเส้นฐาน (baseline) ความลึกคือระยะทางตามแนวดิ่งใต้เส้นฐาน ถึงแม้ว่าจะมีการกำหนดกล่องไว้เช่นนี้แล้วตัวอักษรที่ออกแบบก็ไม่จำเป็นที่จะต้องอยู่ในกล่องเสมอไป อาจจะมีการออกนอกกล่องไปบ้างก็ได้เช่นการออกแบบตัวอักษรที่เป็นตัวเอียงหรือสระที่อยู่ข้างเหนือบรรทัดปกติของภาษาไทย จะมีบางส่วนของตัวอักษรนั้นอยู่นอกกล่อง เป็นต้น

การกำหนดค่ามิติ (dimension) ของกล่องนั้นจะกำหนดในรูปแบบของปริมาณชาร์ป (sharp quantities) กล่าวคือ ขนาดของกล่องจะไม่ขึ้นอยู่กับอุปกรณ์การแสดงผล โปรแกรมเมตาฟอนต์จะเป็นตัวจัดการให้อุปกรณ์นั้นรู้จักกับเอกสารที่สร้างขึ้นโดยการสร้างรูปแบบการทำงาน (mode) ใหม่ แต่ตัวอักษรนั้นจะยังคงมีมิติของกล่องเท่าเดิมเมื่อได้กล่องตามต้องการแล้วก็จะกำหนดพิกเซล (pixel) ในกล่องนั้นการสร้างเวคเตอร์ต่างๆก็จะใช้ค่าของพิกเซลเหล่านี้เป็นจุดอ้างอิง เมื่อทำการสร้างตัวอักษรขึ้นมา1ตัวแล้วก็จะต้องมีการให้ค่าสำหรับตัวอักษรตัวที่สร้างขึ้นนี้ด้วยเพื่อที่เท็กซ์จะได้สามารถอ้างอิงได้ เช่นตัวอักษร "ก" จะแทนที่ด้วยค่า 161 เป็นต้น

หลังจากที่สามารถเขียนโปรแกรมสร้างตัวอักษรได้แล้วก็จะทำการแปลงชุดคำสั่งของโปรแกรมที่เขียนขึ้นให้อยู่ในรูปตัวอักษรแบบทั่วไป (generic font) ซึ่งสามารถเปลี่ยนเป็นตัวอักษรสำหรับอุปกรณ์แสดงผลใดๆก็ได้อย่างไรก็ตามตัวอักษรที่ได้นี้อาจจะยังไม่มีความละเอียดเหมาะสมกับอุปกรณ์แสดงผล ดังนั้นจำเป็นจะต้องอาศัยซอฟท์แวร์เปลี่ยนข้อมูลตัวอักษรแบบทั่วไปเหล่านี้ให้เป็นรูปแบบที่โปรแกรมไดรว์เวอร์ช่วย

ในการแสดงผลสำหรับอุปกรณ์แสดงผลแต่ละชนิดจะรู้จักและนำออกไปสร้างผลลัพธ์ได้ในที่สุด ตัวอักษรที่ได้จากขั้นตอนนี้จะนำไปแสดงผลและใช้อ้างอิงในขั้นตอนต่อไป

2. สร้างโปรแกรมขึ้นมาเพื่ออ่านเอกสารที่เป็นภาษาไทย จัดการแยกระหว่างข้อมูลและคำสั่งของเท็กซ์ แยกข้อมูลที่เป็นภาษาไทยและอังกฤษหรือตัวเลข จากนั้นทำการแทรกสัญญลักษณ์พิเศษ เพื่อบอกการสิ้นสุดคำภาษาไทย หลังจากนั้นเปลี่ยนตัวอักษรภาษาไทยให้อยู่ในรูปแบบ "\cXXX" (สำหรับเท็กซ์ในเวอร์ชันแรกๆ จะไม่สามารถรู้จักตัวรหัสแอสกีโดยตรงต้องมีการใช้คำสั่ง "\charXXX" นำหน้า แต่ในเวอร์ชันต่อๆมาเท็กซ์จะสามารถรู้จักรหัสแอสกีและสามารถนำรหัสนั้นไปค้นหาตัวอักษรในฟอนต์ต่างๆได้เลย) เพื่อให้เข้ากันได้กับตัวอักษรภาษาไทยที่สร้างขึ้นโดยโปรแกรมเมตาฟอนต์ เมื่อสามารถเปลี่ยนให้อยู่ในรูปแบบดังกล่าวได้แล้ว จะทำการล้อมประโยคภาษาไทยนั้นด้วยเครื่องหมาย {\thai...} เพื่อบอกว่ารหัสเหล่านี้ต้องใช้กับฟอนต์ภาษาไทย

เนื่องจากการทำการตัดคำเพื่อแทรกสัญญลักษณ์พิเศษนั้นอัลกอริธึมที่ใช้ยังไม่มีประสิทธิภาพเทียบเท่ากับอัลกอริธึมตัดคำที่ใช้อยู่ในโปรแกรมประมวลผลคำ CU Writer ทำให้รูปประโยคที่ได้ไม่มีความสมบูรณ์เพียงพอ ดังนั้นถ้าสามารถนำอัลกอริธึมตัดคำของ CU Writer มาใช้ก็จะทำให้การตัดคำมีประสิทธิภาพดียิ่งขึ้น

การทำงานของโปรแกรมแบ่งพยางค์ของ CU Writer จะใช้วิธีการค้นหาจุดแบ่งพยางค์ให้ได้มากที่สุดและผิดพลาดน้อยที่สุดด้วยพจนานุกรมที่มีโครงสร้างแบบทรัย (trie) โดยใช้วิธีการสร้างต้นไม้แบ่งคำ (separation tree) สำหรับคำที่จะแบ่ง ทรัย (trie) คือ ต้นไม้ (tree) ชนิดหนึ่งซึ่งจะมองลักษณะของข้อมูลเป็นลำดับ จากเซ็ตของตัวอักษรโดยตรง แทนที่จะมองแต่ละคำเป็นข้อมูลหนึ่งชิ้น ทำให้มีความสามารถจัดการกับข้อมูลที่มีความยาวไม่คงที่ได้ดีและเข้ากันได้ดีกับโปรแกรมที่ทำงานกับอักขระทีละตัวอักษรเช่นโปรแกรมแบ่งคำนี้ นอกจากนี้ทรัยยังมีความสามารถในการจัดการกับคำนำหน้า (prefix) ร่วมของคำในพจนานุกรมโดยการยุบรวมกันหมด ทำให้ลดขนาดของพจนานุกรมลงได้เป็นอันมาก

การเขียนต้นไม้แบ่งคำจะเขียนคำที่ต้องการจะแบ่งทีละ 1 ตัวอักขระจากบนลงล่าง เมื่อถึงจุดที่สามารถแบ่งคำได้ก็จะเกิดการตัดสินใจว่าจะแบ่งหรือไม่ ถ้าตัวอักขระที่เขียนลงไปทางกิ่งด้านขวาของต้นไม้แบ่งคำก็จะยังไม่แบ่งคำและตัวอักขระถัดไปก็ยังเป็นคำเดียวกัน แต่ถ้าไปทางซ้ายมือหมายถึงการแบ่งคำ และตัวอักขระถัดไปจะถือเป็นตัวเริ่มต้นคำใหม่และหมายถึงการได้จุดแบ่งคำเพิ่มขึ้น อีกหนึ่งจุดด้วย

สำหรับอัลกอริธึมในการค้นหาข้อมูลในต้นไม้ของCU Writer จะค้นหาแบบทางลึกก่อน (depth-first search) ร่วมกับการทำงานแบบย้อนกลับ (backtrack) เรียกว่า BWS ในงานวิจัยนี้จะนำเอาอัลกอริธึมนี้มาประยุกต์ให้สามารถล้อมกรอบคำที่ต้องการแบ่งด้วยชุดคำสั่งของเท็กซ์

3. ทำการแทรกชุดคำสั่งของเท็กซ์เข้าไปให้เหมาะสมกับงานที่ต้องการเช่นการจัดระยะระหว่างบรรทัดของภาษาไทย การจัดช่องไปต่างๆ การจัดระยะห่างของช่องว่างของแต่ละย่อหน้า การจัดบรรทัดสำหรับหมายเหตุที่เป็นภาษาไทย สิ่งต่างๆเหล่านี้จะมีความแตกต่างจากการใช้ตัวอักษรที่เป็นภาษาอังกฤษ นอกจากนี้ต้องทำการตรวจสอบสระหรือวรรณยุกต์ที่ต้องมีการปรับระดับให้อยู่เหนือหรือใต้พยัญชนะนั้น แทรกชุดคำสั่งของเท็กซ์เพื่อวางตำแหน่งของสระและวรรณยุกต์เหล่านั้นให้เหมาะสม

หลักการวางตำแหน่งของสระนั้นจะสามารถทำได้ในขั้นตอนของการออกแบบตัวอักษรด้วยโปรแกรมเมตาฟอนต์โดยจะกำหนดให้ตัวอักษรอยู่นอกกรอบไปทางด้านซ้ายมือของจุดเริ่มต้นและกำหนดพิกัดสำหรับตัวอักษรเหล่านั้นด้วยค่าที่เป็นลบ สำหรับในกรณีของสระที่อยู่ข้างบนและมีวรรณยุกต์ ประกอบด้วยนั้นก็จะต้องทำการปรับวรรณยุกต์ให้ไม่ทับซ้อนกับตำแหน่งของสระนั้นด้วย

<u>วัตถุประสงค์ในการทำวิจัย</u>

1. ศึกษาการทำงานของโปรแกรมเท็กซ์
2. ศึกษาวิธีการสร้างตัวอักษรโดยการใช้โปรแกรมเมตาฟอนต์
3. สร้างตัวอักษรไทยโดยใช้โปรแกรมเมตาฟอนต์
4. ออกแบบและพัฒนาโปรแกรมแสดงผลตัวอักษรไทยระบบเท็กซ์

<u>ขอบเขตการทำวิจัย</u>

1. ดัดแปลงอัลกอริธึมตัดคำภาษาไทยของโปรแกรม CU Writer ให้สามารถใช้กับโปรแกรมแสดงผลตัวอักษรไทยบนเท็กซ์ได้

2. โปรแกรมสามารถทำงานบนไมโครคอมพิวเตอร์ที่ใช้ระบบปฏิบัติการ MS-DOS

ขั้นตอนการทำวิจัย

1. ศึกษาการทำงานของโปรแกรมเท็กซ์
2. ศึกษาการทำงานของโปรแกรมเมตาฟอนต์
3. สร้างตัวอักษรภาษาไทยสำหรับใช้ในการทดสอบโปรแกรมไทยเท็กซ์
4. ศึกษาอัลกอริธึมตัดคำภาษาไทยของ CU Writer
5. ศึกษาวิธีการสร้างชุดคำสั่งของเท็กซ์ เพื่อจัดการกับปัญหาของภาษาไทย 3 ระดับ
6. สร้างโปรแกรมเปลี่ยนแฟ้มบรรณาธิกรให้เป็นแฟ้มข้อมูลในรูปแบบของเท็กซ์
7. ทดสอบการใช้งานของโปรแกรมที่สร้างขึ้นโดยใช้โปรแกรมเท็กซ์
8. แก้ไขในรายละเอียดของโปรแกรมที่สร้างขึ้นรวมทั้งตกแต่งตัวอักษรให้สวยงาม
9. ทดสอบอีกครั้งเพื่อความถูกต้อง
10. สรุปผลการวิจัยและจัดทำเอกสารประกอบการวิจัย

ประโยชน์ที่คาดว่าจะได้จากการวิจัย

1. จะได้ฟอนต์ตัวอักษรภาษาไทยที่สร้างขึ้นด้วยโปรแกรมเมตาฟอนต์
2. จะได้โปรแกรมแสดงผลตัวอักษรไทยบนระบบเท็กซ์