ระบบค้นคืนข้อมูลพจนานุกรมข้ามภาษาไทย-อังกฤษ โดยใช้คุณลักษณะของคำพ้องเสียง

Thai-English Cross-Language Transliterated Word Retrieval in a Dictionary

โดย

ธนศักดิ์ ไชยะกุล

รหัส 46066820

อาจารย์ที่ปรึกษา

ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 1 ปีการศึกษา 2548
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	ระบบค้นคืนข้อมูลพจนานุกรมข้ามภาษาไทย-อังกฤษ โดยใช้คุณลักษณะของคำพ้องเสียง
นักศึกษา	นายธนศักดิ์ ไชยะกุล
อาจารย์ที่ปรึกษา	ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

โครงการพัฒนาระบบนี้ได้ประยุกต์วิธีการเข้ารหัสคำทับศัพท์และค้นคืนสำหรับระบบค้นคืนข้อมูลพจนานุกรมข้ามภาษาไทย-อังกฤษ โดยใช้คุณลักษณะของคำพ้องเสียง โดยอาศัยหลักการการเข้ารหัสคำทับศัพท์ทั้งคำไทยทับศัพท์คำอังกฤษ และคำอังกฤษทับศัพท์คำไทย สำหรับการเข้ารหัสคำไทยนั้นจะประกอบด้วยขั้นตอนการตัดวรรณยุกต์และไม้ไต่คู้ การเปลี่ยนรูปการสะกดและการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล ส่วนคำในภาษาอังกฤษจะถูกเข้ารหัสโดยการแปลงพยัญชนะอังกฤษเป็นอักษรไทย และการแปลงสระในภาษาอังกฤษให้เป็นสระในภาษาไทย คำที่เข้ารหัสแล้วจากข้อคำถามจะถูกเปรียบเทียบกับรหัสเสียงของคำศัพท์แต่ละคำในฐานข้อมูลพจนานุกรม โดยใช้เทคนิค N-Gram เพื่อวัดค่าความเหมือนระหว่างรหัสเสียงของคำทั้งสอง จากการทดสอบพบว่าระบบมีค่าความแม่นยำและค่าเรียกคืนเฉลี่ยประมาณ 60-70% โครงการพัฒนาระบบนี้พัฒนาขึ้นในรูปแบบของเว็บแอพพลิเคชั่นโดยใช้ ASP.NET ทำให้สามารถให้บริการพจนานุกรมแก่ผู้ที่สนใจจากที่ต่างๆ ทั่วโลกผ่านทางเครือข่ายใยแมงมุม

Title	Thai-English Cross-Language Transliterated Word Retrieval in a Dictionary
Student	Mr. Thanasak Chaiyakul
Advisor	Asst. Prof. Dr. Pattarachai Lalitrojwong
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

This project applys transliterated word encoding and retrieval algorithms for Thai-English cross-language transliterated word retrieval in a dictionary. The encoding is used for both Thai-to-English and English-to-Thai transliterated words. The encoding algorithm transforms Thai words into a phonetical form using romanization rules and transforms English characters to Thai using transliteration rules. The retrieval algorithm searches codes of the query words approximately with codes in the dictionary database using N-Grams based techniques. Experimental results showed that the system can achieve both precision and recall as high as 60-70% simultaneously. This project focuses on developing a Thai-English dictionary system as a web-based application by using ASP.NET. The result will provide dictionary service to users around the world via WWW.

กิตติกรรมประกาศ

กระผมขอขอบคุณบุคคลต่าง ๆ ที่ให้ความช่วยเหลือและส่งเสริมจนการพัฒนาโครงการนี้สำเร็จลงได้ด้วยดี ดังต่อไปนี้

- ขอขอบคุณ มารดาของกระผม ที่สนับสนุนด้านทุนการศึกษาเล่าเรียน
- ขอขอบคุณ ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์ อาจารย์ที่ปรึกษา สำหรับคำปรึกษาและคำแนะนำที่เป็นประโยชน์ จนการพัฒนาโครงการนี้สำเร็จลุล่วงไปด้วยดี
- ขอขอบคุณ ภรรยาของกระผม สำหรับกำลังใจและแรงเชียร์
- ขอขอบคุณ เพื่อน ๆ สำหรับมิตรภาพดี ๆ ที่มีให้กันเสมอมา
- ขอขอบคุณ คณาจารย์และเจ้าหน้าที่ทุกท่านที่ให้ความรู้และความช่วยเหลือต่าง ๆ ที่กระผมได้รับตลอดระยะเวลาที่ศึกษาอยู่ ณ สถาบันแห่งนี้
- ขอขอบคุณ ที่ทำงานของกระผม บริษัท ดีเอสที อินเทอร์เนชั่นเนล จำกัด ที่เปิดโอกาสให้ผมออกจากที่ทำงานก่อนเวลาเลิกงาน ในวันที่กระผมมีเรียน

นายธนศักดิ์ ไชยะกุล

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูป

หน้า

รูปที่

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มา

ปัจจุบันการใช้ภาษาอังกฤษได้เข้ามาเป็นส่วนสำคัญในชีวิตประจำวันของคนไทยมากขึ้น ไม่ว่าจะใช้ในการทำงาน การเรียน หรือการท่องเที่ยวต่างประเทศ ล้วนแล้วแต่มีการใช้ภาษาอังกฤษมาเกี่ยวข้องด้วยทั้งสิ้น พจนานุกรมจึงเข้ามามีบทบาทเป็นอย่างมาก เพื่อใช้ค้นหาความหมายของคำศัพท์ที่ต้องการ แต่พบว่าพจนานุกรม ภาษาไทย-อังกฤษ โดยทั่วไปนั้น ไม่สามารถใช้ในการค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่แสดงในการสอบถามได้ ทัศนวรรณ ศูนย์กลาง และคณะ (2543 : 159) ได้กล่าวถึงปัญหาในการค้นคืนคำศัพท์ข้ามภาษาซึ่งคำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลายรูปแบบ เช่น “Carbohydrate” ในภาษาไทยอาจพบได้ทั้ง “คาร์โบไฮเดรต” “คาร์โบไฮเดรท” หรือ “คาร์โบฮัยเดรต” การนำพจนานุกรมสองภาษา (Bilingual Dictionary) มาใช้ในระบบค้นคืนคำศัพท์และความหมายจากพจนานุกรมไม่อาจแก้ปัญหานี้ได้มากนัก เนื่องจากคำทับศัพท์ส่วนมากมักไม่ปรากฏในพจนานุกรม ซึ่งระบบค้นคืนทั่วไปที่มีอยู่ไม่สามารถแก้ปัญหาเหล่านี้ได้ รวมถึงปัญหาเกี่ยวกับการจดจำคำศัพท์ต่าง ๆ ทำให้บ่อยครั้งทำให้สะกดคำศัพท์ผิด เป็นต้น แต่ส่วนใหญ่ผู้ใช้งานพจนานุกรมมักจะสามารถจดจำคำอ่านได้ ด้วยเหตุนี้จึงได้เกิดแนวคิดที่จะพัฒนาระบบค้นคืนข้อมูลคำศัพท์จากพจนานุกรมโดยใช้คุณลักษณะคำพ้องเสียง ซึ่งรวมถึงการค้นคืนข้อมูลคำศัพท์คำอ่าน และคำทับศัพท์ด้วย เพื่อเพิ่มความสามารถให้ระบบพจนานุกรมไทย-อังกฤษ ให้สามารถค้นหาคำที่อ่านออกเสียงคล้ายกัน ซึ่งเหมาะสำหรับการค้นหาคำศัพท์ที่ไม่แน่ใจในตัวสะกด และเพื่อเพิ่มความสามารถในการให้บริการค้นหาคำศัพท์จากพจนานุกรมที่มีอยู่ในปัจจุบัน เช่นกรณีที่ผู้ใช้งานป้อนคำหลักด้วยคำใดคำหนึ่งในภาษาใดภาษาหนึ่ง ในขณะที่คำหลักในเอกสารจัดเก็บด้วยภาษาอื่น หรือภาษาเดียวกันที่สะกดไม่เหมือนกันแต่ออกเสียงคล้ายกัน เช่น ถ้าต้องการค้นหาคำว่า "ด๊อกเตอร์" ระบบจะต้องสามารถค้นคืนคำศัพท์ที่สะกดว่า "Doctor" กลับมายังผู้ค้นหา เป็นต้น

อีกทั้งปัจจุบันมีการใช้งานเครือข่ายอินเทอร์เน็ตและเทคโนโลยีสารสนเทศอย่างแพร่หลาย หนึ่งในการให้บริการที่ได้รับความนิยมเป็นอย่างสูงได้แก่ การใช้บริการผ่านเว็บ หรือโฮมเพจ หนึ่งในบริการเหล่านั้นได้แก่ บริการค้นหาคำศัพท์จากพจนานุกรมผ่านเว็บ เช่น http://lexitron.nectec.or.thโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ซึ่งให้บริการค้นหาคำศัพท์ภาษาไทย-อังกฤษผ่านเว็บ และ http://www.sansarn.com โดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ที่ให้บริการสืบค้นโดยใช้ซาวเด็กซ์ภาษาไทย

1.2 วัตถุประสงค์ของโครงการ

ในการพัฒนาระบบค้นคืนข้อมูลพจนานุกรมข้ามภาษาไทย-อังกฤษ โดยใช้คุณลักษณะของคำพ้องเสียง มีวัตถุประสงค์ในการพัฒนาระบบงานดังนี้

1. เพื่อศึกษาการเข้ารหัสคำศัพท์ภาษาไทยให้เป็นรหัสเสียงโดยใช้อักษรแทนเสียง
2. เพื่อศึกษาการเข้ารหัสคำศัพท์ภาษาอังกฤษให้เป็นรหัสเสียงโดยใช้อักษรแทนเสียง
3. เพื่อศึกษาการเข้ารหัสคำศัพท์ในฐานข้อมูลพจนานุกรมไทย-อังกฤษ
4. เพื่อเพิ่มความสามารถให้ระบบสามารถค้นหาคำศัพท์ข้ามภาษาโดยใช้คำพ้องเสียงหรือคำอ่าน
5. เพื่อเพิ่มความสามารถค้นหาคำศัพท์ในหลายรูปแบบต่าง ๆ ได้แก่ ไทย-ไทย ไทย-อังกฤษ อังกฤษ-ไทย และ อังกฤษ-อังกฤษ
6. เพื่อลดปัญหาในการจดจำตัวสะกดของคำศัพท์
7. เพื่อนำเทคโนโลยีอินเทอร์เน็ตมาประยุกต์ใช้กับระบบฐานข้อมูลพจนานุกรม เพื่อเพิ่มช่องทางการให้บริการแก่บุคคลทั่วไป

1.3 ขอบเขตการศึกษา

1. ศึกษาอัลกอริทึมในการเข้ารหัสคำศัพท์ภาษาไทย และภาษาอังกฤษ
2. ศึกษาอัลกอริทึมในการเปรียบเทียบค่าความแตกต่างของรหัสคำ
3. ศึกษาการรายงาน วารสาร และเอกสารวิชาการต่าง ๆ ที่เกี่ยวข้อง
4. ศึกษาเทคโนโลยีที่ใช้พัฒนาระบบงาน ทั้งส่วนของฮาร์ดแวร์และซอฟต์แวร์
5. ศึกษาโครงสร้างแฟ้มข้อมูลของพจนานุกรม LEXiTRON ที่จะนำมาจัดเก็บในฐานข้อมูลของระบบ
6. ศึกษาการพัฒนาระบบโดยใช้โปรแกรม Microsoft Visual Studio .NET 2005 และ Microsoft SQL Server 2005
7. ศึกษาความรู้เกี่ยวกับเอกสาร XML

1.4 ขั้นตอนการพัฒนาระบบ

โครงการพัฒนาระบบงานประกอบด้วยขั้นตอนการพัฒนาระบบ ดังต่อไปนี้

1. ศึกษาการเข้ารหัสคำทั้งภาษาไทยและอังกฤษให้เป็นรหัสเสียงซึ่งเปรียบได้กับภาษากลาง
2. ศึกษาเทคนิคการนำคำศัพท์ที่เข้ารหัสแล้วมาเปรียบเทียบกับข้อมูลจากพจนานุกรม ซึ่งคำในพจนานุกรมจะถูกนำมาเข้ารหัสเช่นเดียวกัน

3. ศึกษาเทคนิคการคำนวณหาค่าความแตกต่างของรหัสคำ จะได้มาจากการคำนวณหาค่าความแตกต่างด้วยวิธี N-Gram ซึ่งจะพิจารณาความแตกต่างกันของอักขระแทนเสียง โดยใช้กลุ่มอักขระของซาวเด็กซ์ช่วยในการกำหนดกลุ่มเสียงที่คล้ายกัน
4. ศึกษาวิธีการการคัดเลือกผลลัพธ์ที่ได้จะเป็นชุดคำศัพท์ที่ผ่านเกณฑ์ (Threshold) ที่ตั้งไว้ พร้อมทั้งแสดงความหมายของคำศัพท์เหล่านั้น
5. ศึกษาการโอนถ่ายฐานข้อมูลคำศัพท์จากพจนานุกรม LEXiTRON มายังฐานข้อมูลพจนานุกรมของระบบ
6. ศึกษาวิธีการค้นหาคำศัพท์ตามคำพ้องเสียงในหลายรูปแบบต่าง ๆ ได้แก่ คำไทย-คำไทย คำไทย-คำอังกฤษ คำอังกฤษ-คำไทย และคำอังกฤษ-คำอังกฤษ
7. สรุปผลการศึกษาและเสนอแนะข้อคิดเห็น จากการศึกษาและพัฒนาระบบ ตลอดจนจัดทำเอกสารการพัฒนาระบบ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เป็นต้นแบบในการค้นคืนข้ามภาษาโดยใช้คำพ้องเสียงหรือคำอ่าน และสามารถประยุกต์ใช้กับระบบอื่นๆ ได้ในอนาคต
2. สามารถค้นหาคำศัพท์ตามคำพ้องเสียงในหลายรูปแบบต่าง ๆ ได้แก่ ไทย-ไทย ไทย-อังกฤษ อังกฤษ-ไทย และ อังกฤษ-อังกฤษ
3. ลดปัญหาในการจดจำตัวสะกดของคำศัพท์
4. เพิ่มช่องทางการให้บริการแก่บุคคลทั่วไป โดยนำเทคโนโลยีเว็บ บนเครือข่ายอินเทอร์เน็ตมาประยุกต์ใช้กับระบบฐานข้อมูลพจนานุกรม
5. ระบบสามารถค้นหาคำแปลตามความหมาย ได้เช่นเดียวกับพจนานุกรม ไทย-อังกฤษ อังกฤษ-ไทย ทั่วไป

บทที่ 2
เทคโนโลยีและทฤษฎีที่เกี่ยวข้อง

2.1 การค้นคืนสารสนเทศ

ระบบค้นคืนสารสนเทศเป็นเครื่องมือที่สำคัญอย่างยิ่งในการบริหารสารสนเทศที่มีอยู่จำนวนมาก โดยปกติแล้วการวัดประสิทธิผลการค้นคืนสารสนเทศใด ๆ มักจะวัดจากค่าความแม่นยำ (Precision) และ ค่าเรียกคืน (Recall)

2.1.1 ค่าแม่นยำ

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542 : 1) อธิบายความหมายของค่าแม่นยำ ว่าหมายถึง "การวัดความสามารถของระบบในการขจัดสารสนเทศที่ไม่เกี่ยวข้องออกไป ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของสารสนเทศที่เกี่ยวข้องที่คืนกลับมา กับจำนวนสารสนเทศทั้งหมดที่กลับคืนมา"

2.1.2 ค่าเรียกคืน

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542 : 1) อธิบายความหมายของค่าเรียกคืน ว่าหมายถึง "การวัดความสามารถของระบบในการคืนสารสนเทศที่เกี่ยวข้องกลับมา ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของสารสนเทศที่เกี่ยวข้องที่คืนกลับมา กับจำนวนสารสนเทศทั้งหมดที่เกี่ยวข้อง"

ในกรณีที่ผู้ใช้ป้อนคำศัพท์ด้วยภาษาใดภาษาหนึ่ง ในขณะที่คำศัพท์ที่จัดเก็บด้วยภาษาอื่น ตัวอย่างเช่น ผู้ใช้ต้องการค้นหาคำที่ออกเสียงว่า "พาติชั่น" แต่ระบบไม่ได้คืนคำศัพท์ว่า "Partition" (คำทับศัพท์ที่ตรงกัน) ทำให้ค่าเรียกคืนของระบบค้นคืนสารสนเทศน้อยกว่าที่ควรจะเป็น ถ้าระบบดังกล่าวไม่สนับสนุนการทำงานแบบข้ามภาษา (Cross-Language Retrieval)

2.2 การค้นคืนข้ามภาษา

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542 : 1) อธิบายความหมายของการค้นคืนข้ามภาษาว่าหมายถึง "การค้นคืนสารสนเทศ โดยภาษาที่ใช้ในข้อคำถามแตกต่างจากภาษาที่ใช้ในการจัดเก็บเอกสาร การใช้พจนานุกรมสองภาษา (Bilingual Dictionary) ในลักษณะของอรรถาภิธาน (Thesaurus) กับระบบค้นคืนสารสนเทศไม่สามารถแก้ไขปัญหาดังกล่าวได้มากนัก เนื่องจากคำทับศัพท์ส่วนมากมักเป็นคำเฉพาะที่ไม่ปรากฏในพจนานุกรม" โดยเฉพาะใน

ปัจจุบันสารสนเทศในสื่ออิเล็กทรอนิกส์มีจำนวนเพิ่มขึ้นอย่างรวดเร็วมาก ทำให้ปัญหาดังกล่าวยิ่งเพิ่มมากขึ้นจนระบบค้นคืนทั่วไปที่มีอยู่ไม่สามารถแก้ปัญหาได้

2.3 การเข้ารหัสคำศัพท์ภาษาไทยและคำศัพท์ภาษาอังกฤษ

ในการเปรียบคำเดียวกันแต่จัดเก็บคนละภาษาจำเป็นจะต้องมีภาษากลางเพื่อที่จะสามารถเปรียบเทียบความแตกต่างของคำทั้งสองได้ โดยการใช้อักษรแทนรหัสเสียง ซึ่งขั้นตอนวิธีการเข้ารหัสคำแบ่งออกเป็นสองกระบวนการ ได้แก่

1. การเข้ารหัสคำสำหรับคำไทยทับศัพท์คำอังกฤษ โดยอาศัยหลักการถ่ายเสียง
2. การเข้ารหัสคำสำหรับคำอังกฤษทับศัพท์คำไทย โดยอาศัยหลักการแปลงอักษร

การเข้ารหัสคำมีจุดประสงค์เพื่อแปลงคำไทยทับศัพท์คำอังกฤษ และคำอังกฤษที่ทับศัพท์คำไทย ให้อยู่ในรูปรหัสคำรูปแบบเดียวกัน รหัสคำเฉพาะทั้งคำไทย และคำอังกฤษที่ทับศัพท์คำไทยต่าง ๆ ในเอกสาร จะถูกสร้างขึ้นในขั้นตอนการสร้างดัชนีในระบบการจัดเก็บสารสนเทศ เมื่อผู้ใช้ป้อนข้อคำถามที่เป็นคำเฉพาะคำไทย หรือคำอังกฤษทับศัพท์คำไทย ระบบค้นคืนก็จะสร้างรหัสคำของคำต่าง ๆ ในข้อคำถาม เพื่อใช้ค้นหากับรหัสคำที่จัดเก็บไว้ในดัชนีของพจนานุกรม

2.4 การเข้ารหัสคำสำหรับคำไทยทับศัพท์คำอังกฤษ

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542 : 1-2) อธิบายถึงขั้นตอนการเข้ารหัสคำไทยที่ทับศัพท์คำอังกฤษไว้ดังนี้ "ขั้นตอนแปลงรูปคำไทย มีจุดประสงค์หลักเพื่อแปลงคำไทยที่อ่านออกเสียงคล้ายกัน แต่เขียนได้หลายรูปแบบ ให้อยู่ในรูปแบบเดียวกัน เพื่อให้ขั้นตอนการเทียบรหัสกระทำได้ง่ายขึ้น การแปลงรูปประกอบด้วยการตัดวรรณยุกต์และไม้ไต่คู้ การเปลี่ยนรูปการสะกด และการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล"

2.4.1 การตัดวรรณยุกต์และไม้ไต่คู้

เนื่องจากในการแปลงอักษรไทยเป็นอักษรอังกฤษนั้น จะไม่พิจารณาวรรณยุกต์และไม้ไต่คู้ เช่น ช้าง แปลงอักษรเป็น Chang ดังนั้น วรรณยุกต์ และไม้ไต่คู้ จะถูกตัดทิ้งจากคำไทย

2.4.2 การเปลี่ยนรูปการสะกด

การเปลี่ยนรูปการสะกดอาศัยหลักทางภาษาศาสตร์ และข้อมูลทางสถิติ เพื่อเปลี่ยนรูปแบบการสะกดของคำไทยที่เขียนได้หลายรูปแบบให้เหมือนกัน แบ่งเป็นกรณีต่าง ๆ ดังนี้

- <u>รร</u> ทำการเปลี่ยน รร เป็น อัน ในกรณีที่ไม่มีตัวสะกดตามหลัง เช่น จรรยา บรรจบ ครรภ์ เป็นต้น และ –ั ในกรณีที่มีตัวสะกดตามหลัง เช่น ธรรม พรรณ กรรม เป็นต้น
- <u>สระ ใ- ไ- และ ไ-ย</u> ทำการเปลี่ยนสระดังกล่าวทั้งหมดเป็น -ย

- <u>สระ ำ และ ัม</u> ทำการเปลี่ยนสระ ำ เป็น ัม
- <u>การันต์และอักษรควบการันต์</u> ทำการตัดอักษรและอักษรควบที่มีตัวการันต์กำกับออก เนื่องจากการแปลงอักษรโดยปกติจะไม่แปลงอักษรที่มีการันต์กำกับ เช่น สิทธิ์ แปลงเป็น Sith พันธุ์ แปลงเป็น Phan เป็นต้น

2.4.3 การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล

การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากลมีจุดประสงค์เพื่อแปลงทำให้คำไทยที่มีการใช้สระประสมและสระเดี่ยวที่ใช้อักขระตั้งแต่สองตัวขึ้นไป ให้อยู่ในรูปแบบที่ง่ายต่อการประมวลผล และลดจำนวนอักขระที่ต้องเปรียบเทียบ ซึ่งขั้นตอนนี้จะใช้เสียงสากล แทนเสียงสระดังกล่าว โดยจะวางสัญลักษณ์เสียงสากลไว้หลังพยัญชนะต้น เพื่อให้มีรูปแบบเหมือนกับผลลัพธ์ที่ได้จากการแปลงอักษรคำอังกฤษทับศัพท์คำไทย การใช้สัญลักษณ์เสียงสากลแทนที่สระมีดังนี้

- ใช้สัญลักษณ์เสียง e แทนสระ เ-ะ เช่น เละ เปลี่ยนเป็น *le*
- ใช้สัญลักษณ์เสียง x แทนสระ แ-ะ แ- -ะ เช่น แกะ แหละ เปลี่ยนเป็น *kx hlx* ตามลำดับ
- ใช้สัญลักษณ์เสียง q แทนสระ เ-ิ เ- ิ เช่น เกิด เผชิญ เปลี่ยนเป็น *kqd phchqy* ตามลำดับ
- ใช้สัญลักษณ์เสียง I แทนสระ เ-ียะ เ- ียะ เ-ีย และ เ- ีย เช่น เสียง เกวียน เปลี่ยนเป็น sIng kwIn ตามลำดับ
- ใช้สัญลักษณ์เสียง U แทนสระ เ-ือะ เ-ือ เ- ือ -ัวะ และ -ัว เช่น เรือง เกลือ ผัวะ และ วัว เปลี่ยนเป็น r Ung klU phU และ wU ตามลำดับ
- ใช้สัญลักษณ์เสียง @ แทนสระ เ-า เ- -า เ-าะ และ เ- -าะ เช่น เงา เพรา และ เฉพาะ เปลี่ยนเป็น ง@ พร@ ง@ และ ฉพ@ ตามลำดับ

2.5 การเข้ารหัสคำสำหรับคำอังกฤษทับศัพท์คำไทย

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542 : 2-3) อธิบายเพิ่มเติมในกรณีที่ข้อคำถามเป็นคำอังกฤษทับศัพท์คำไทย จะทำการแปลงอักษรอังกฤษเป็นไทย การแปลงอักษรที่นำเสนอในที่นี้เป็นการเปลี่ยนพยัญชนะอังกฤษเป็นไทย ส่วนสระอังกฤษจะใช้หลักเกณฑ์การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากลคือจะแปลงสระอังกฤษเป็นสระไทย แต่ถ้าสระไทยนั้นเป็นสระที่ใช้อักขระตั้งแต่สองตัวขึ้นไป จะใช้สัญลักษณ์เสียงสากล แทนเสียงสระดังกล่าว

หลักเกณฑ์ในการแปลงอักษรในส่วนพยัญชนะ จะใช้หลักการเทียบตัวอักษรโรมัน-ไทยของ ISO โดยได้ดัดแปลงบางส่วนเพื่อให้สมบูรณ์ยิ่งขึ้น เช่น แปลง DH เป็น ท และแปลง BH เป็น พ เนื่องจากเป็นการเขียนแบบบาลีสันสกฤต ซึ่งยังมีใช้กันอยู่มากในปัจจุบัน หลักเกณฑ์ในการแปลงอักษรในส่วนพยัญชนะแสดงไว้ดังตารางที่ 2.1

ตารางที่ 2.1 การแปลงอักษรพยัญชนะอังกฤษเป็นไทย (ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล. 2542 : 3)

อังกฤษ	ไทย		อังกฤษ	ไทย	
B	บ		N	น (ณ)	
BH	พ	*	NG	ง	
C	ซ	*	P	ป	
CH	ช (ฉ ฌ)		PH	พ (ผ ภ)	
CK	ก	*	Q	ค	*
D	ด (ฎ)		R	ร (ฤ)	
DH	ท	*	S	ส (ซ ศ ษ)	
F	ฟ (ฝ)		T	ต (ฏ)	
G	ก	*	TH	ท (ฐ ฑ ฒ ถ ธ)	
H	ห ฮ		V	ว	
J	จ	*	W	ว	
K	ก		X	ก	*
KH	ข (ฃ ค ฅ ฆ)		Y	ย (ญ)	
L	ล (ฦ ฬ)		Z	ซ	*
M	ม				

* = ส่วนที่ปรับเปลี่ยนและเพิ่มเติมจากแบบของ ISO

ส่วนหลักเกณฑ์ในการแปลงอักษรในส่วนสระอังกฤษเป็นสระไทยนั้น ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (2542) พบว่ามีปัญหาอย่างมากในการแปลงอักษร คือ หนึ่งหน่วยอักษรโรมันสามารถแปลงได้เป็นหลายหน่วยอักษรไทย เช่น A แปลงอักษรเป็น เ- แ-ะ -า และ O แปลงอักษรเป็น –อ โ- เป็นต้น และหลายหน่วยอักษรโรมัน สามารถแปลงเป็นหนึ่งหน่วยอักษรไทยได้ เช่น U หรือ OO แปลงเป็นสระ -ุ เป็นต้น และจากการศึกษาของผู้วิจัยพบว่ามีความ

หลากหลายในการใช้อักษรโรมันแทนภาษาไทย เช่น คำว่า พร มีการเขียนเป็น Phon, Phorn, Porn, Pon เป็นต้น ดังนั้น จึงได้พยายามยึดหลักการแปลงอักษรของราชบัณฑิตยสถาน และการใช้อักษรโรมันแทนอักษรไทยของจันทร์เพ็ญ โวหารสุนทร (2530) ได้จากการสำรวจความนิยมในการใช้อักษรโรมันแทนอักษรไทย เป็นต้นแบบและเพิ่มเติมบางส่วนจากการศึกษาของผู้วิจัยที่มีผู้นิยมใช้เข้าไปดังตารางที่ 2.1

ถ้าตัวอักษรแรกของคำเป็นสระได้แก่ A, E, I, O และ U ให้แปลง A เป็น อ แปลง E เป็น เอ (สระ เ- กับ อ) แปลง I เป็น อิ (อ กับสระ ิ) แปลง O เป็น โอ (สระ โ- กับ อ) และแปลง U เป็น อุ (อ กับสระ ุ) และถ้าตัวอักษรถัดไปเป็นสระอีก ให้นำอักษรตัวแรกไปรวมด้วยในการแปลงอักษรโดยเทียบตามตารางที่ 2.2

ตารางที่ 2.2 การแปลงอักษรสระอังกฤษเป็นไทย (ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล. 2542 : 4)

อังกฤษ	ถอดอักษรเป็น	หมายเหตุ
A	ะ	
AA	า	*
AE	x (แ-ะ แ-)	
AI	ัย	
AO	@ (เ-า)	
AIU	I (เ ียว)	*
ARN	าน	*
ART	าท	*
E	เ (เ-ะ เ-)	
EE	ี	*
EO	แ-ว	
ER	q (เ-อ เ ิ) หลัง R ต้องไม่เป็นสระ	
EU	ี	

* = ส่วนที่เพิ่มเติมจากแบบของราชบัณฑิตยสถาน

ตารางที่ 2.2 การแปลงอักษรสระอังกฤษเป็นไทย (ต่อ)

อังกฤษ	ถอดอักษรเป็น	หมายเหตุ
I	ิ	
IA	I (เ ียะ เ ีย)	
IE	I (เ ียะ เ ีย)	*
O	อ (-อ)	
OE	ๆ (เ-อ เ ิ)	
OI	อย	
OO	ู	
ORN	ร (-อน)	*
U	ุ (ึ ื ุ ู)	
UA	U (เ ือะ เ ือ ัวะ ัว)	
UE	ึ	

2.6 การเปรียบเทียบรหัสคำและการค้นคืน

รหัสคำที่ได้ของคู่คำไทยและคำภาษาอังกฤษทับศัพท์คำไทยนั้นอาจไม่ตรงกันทุกตัวอักษร แต่จะมีลักษณะคล้ายกัน ทั้งนี้ เนื่องจากหลักการทับศัพท์ที่ใช้กันในปัจจุบันมีหลายรูปแบบ การเปรียบเทียบรหัสคำแบบเปรียบเทียบตัวต่อตัวในรหัสให้ตรงกัน การเทียบรหัสคำแบบประมาณโดยอาศัยการคำนวณค่าความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคที่เรียกว่า N-Gram Based Techniques จากนั้น นำค่าความแตกต่างของคู่รหัสคำที่ได้มาทดสอบกับเงื่อนไขในการเปรียบเทียบ ถ้าผ่านการทดสอบจะสรุปได้ว่ารหัสคำทั้งสองรหัสเป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา

2.6.1 N-Grams Based Techniques

Holmes and McCabe (2002) ได้อธิบายถึงเทคนิคในการเปรียบเทียบความคล้ายกันของคำทั้งสอง โดยจะคำนวนหาค่า δ จากการหารจำนวนตัวอักษรที่ซ้ำกันของคำทั้งสอง เรียกว่า Common N-Grams ด้วยจำนวนตัวอักษรทั้งหมดที่ไม่ซ้ำกันของคำทั้งสอง สมมติให้ A และ B คือคำที่นำมาเปรียบเทียบกัน เราจะสามารถหาค่า δ จากสูตรต่อไปนี้ โดย δ จะมีค่าตั้งแต่ 0 ถึง 1 โดยถ้า δ มีค่าเท่า 1 สามารถสรุปได้ว่าคำทั้งสองเหมือนกันทุกประการ แสดงสมการไว้ดังรูปที่ 2.1

$$\delta = A\Omega B / A U B$$

รูปที่ 2.1 สมการคำนวนหาค่าความคล้ายคลึงระหว่างคำทั้งสองโดยใช้ N-Grams Based Techniques

2.6.2 เกณฑ์การเปรียบเทียบคำแบบประมาณ

จากการเปรียบเทียบความเหมือนของรหัสคำด้วย N-Grams Based Techniques เพื่อให้ง่ายต่อการกำหนดค่าความเหมือนที่ยอมรับได้จากผู้ใช้ระบบ ผู้พัฒนาระบบจึงได้ปรับเปลี่ยนค่าของ δ ให้มีหน่วยเป็นเปอร์เซ็นต์ โดยรหัสคำของ A และ B ใด ๆ ที่มีความแตกต่างผ่านเกณฑ์ที่แสดงข้างล่างนี้จะถือว่าเป็นรหัสคำของคำที่ตรงกันระหว่างภาษาไทยกับภาษาอังกฤษทับศัพท์ภาษาไทย เกณฑ์การเปรียบเทียบคำแบบประมาณ แสดงไว้ดังรูปที่ 2.2

$$(\delta \times 100) >= \alpha$$

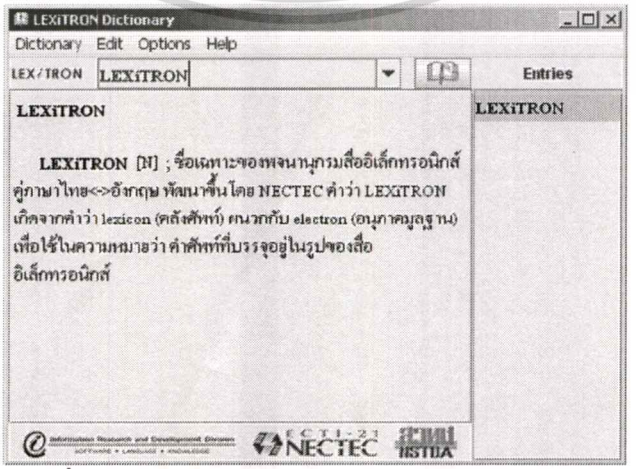
รูปที่ 2.2 เกณฑ์การเปรียบเทียบคำแบบประมาณ

โดย δ x 100 หมายถึง เปอร์เซ็นต์ความเหมือนระหว่างรหัสคำของ A และ B มีค่าอยู่ในช่วง 0% (แตกต่างกันโดยสิ้นเชิง) ถึง 100% (เหมือนกันทุกประการ) ในขณะที่ α นี้เป็นพารามิเตอร์ของระบบที่ผู้ใช้สามารถกำหนดได้ ซึ่งจะส่งผลต่อค่าแม่นยำ และค่าเรียกคืนของระบบ α มีค่าระหว่าง 0 ถึง 100 เช่นกัน ซึ่งจะเป็นตัวกำหนดเกณฑ์การยอมรับความเหมือนกันของรหัสคำแบบประมาณ โดยที่ 100 หมายถึงรหัสคำต้องเหมือนกันทุกประการจึงจะยอมรับ ในขณะที่ 0 หมายถึงการยอมรับทุก ๆ คู่รหัสคำไม่ว่าจะแตกต่างกันเท่าใดก็ตาม

บทที่ 3
เครื่องมือที่ใช้พัฒนาโครงการ

3.1 พจนานุกรม LEXiTRON

พจนานุกรม LEXiTRON เป็นพจนานุกรมสื่ออิเล็กทรอนิกส์ ไทย - อังกฤษ พัฒนาขึ้นโดย ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ หลักการที่สำคัญของการพัฒนา พจนานุกรมนี้คือ ความต้องการที่จะนำเทคโนโลยีฐานข้อมูลขนาดใหญ่เข้ามาช่วยในการวิจัยและพัฒนาในสาขาการประมวลผลภาษาธรรมชาติ ด้วยแนวคิดที่ว่า ฐานข้อมูลขนาดใหญ่สามารถสะท้อนรูปแบบของภาษาโดยรวมได้ เราเรียกการพัฒนาฐานข้อมูลพจนานุกรมในลักษณะนี้ว่า การสร้างพจนานุกรมจากฐานข้อมูลขนาดใหญ่ (Corpus-based dictionary) ซึ่งหลักการของการสร้างฐานข้อมูลชนิดนี้คือ การรวบรวมคำศัพท์จากคำที่ปรากฏใช้จริงในอัตราความถี่สูงในบริบทต่างๆ ของการใช้ภาษา โดยใช้เทคโนโลยีทางคอมพิวเตอร์รวบรวมและคัดเลือกมาจากฐานข้อมูลขนาดใหญ่ จากแนวคิดดังกล่าว ทำให้ LEXiTRON เป็นพจนานุกรมค้นหาศัพท์คู่ภาษา ไทย อังกฤษ ที่ประกอบด้วยศัพท์ทันสมัยที่มีใช้อยู่ในปัจจุบันจำนวนมาก และนอกเหนือจากคุณประโยชน์ที่ได้จากการใช้เทคโนโลยีทางด้านคอมพิวเตอร์ผสมผสานกับภาษาศาสตร์เพื่อประโยชน์ทางการใช้ที่ง่ายและรวดเร็วขึ้นนั้น ยังถือเป็นประโยชน์อย่างยิ่งสำหรับการค้นคว้าข้อมูลเกี่ยวกับคำศัพท์เพื่อใช้ประโยชน์ทางด้านการวิจัยทางภาษาอย่างยิ่ง (ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 2546) ตัวอย่างหน้าจอของโปรแกรมพจนานุกรม LEXiTRON แสดงไว้ดังรูปที่ 3.1



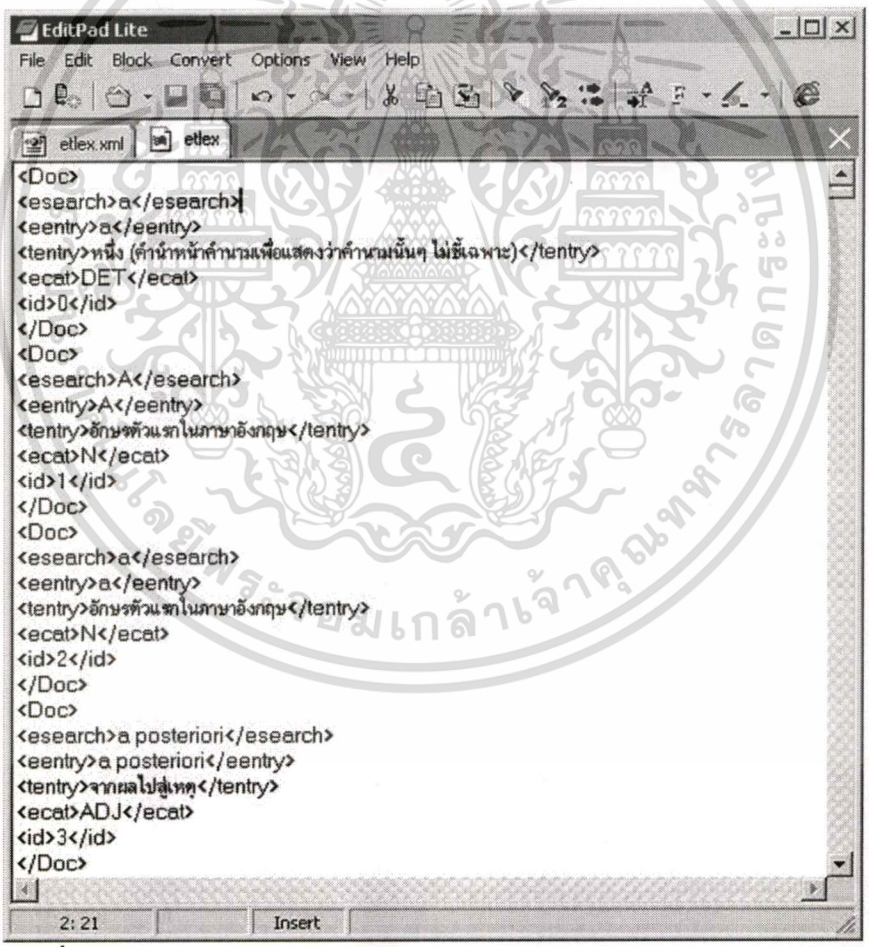
รูปที่ 3.1 หน้าจอโปรแกรมพจนานุกรม LEXiTRON

3.1.1 โครงสร้างของแฟ้มข้อมูล LEXiTRON

โครงสร้างแฟ้มข้อมูลของ LEXiTRON มีลักษณะคล้ายกับเอกสาร XML แต่ยังมีรูปแบบไม่ถูกต้อง ดังนั้นจึงจำเป็นต้องปรับปรุงโครงสร้างและข้อมูลให้อยู่ในรูปแบบของ XML ที่เหมาะสมเสียก่อน จึงจะสามารถนำไปใช้ประโยชน์ต่อไปได้ โดยการปรับปรุงมีสองขั้นตอนได้แก่

1. เพิ่มแท็ก <Docs> … </Docs> ซึ่งเป็นรูทอิลิเมนต์ให้แก่เอกสาร XML
2. การตัดอักขระที่สงวนไว้ออกจากเอกสาร XML

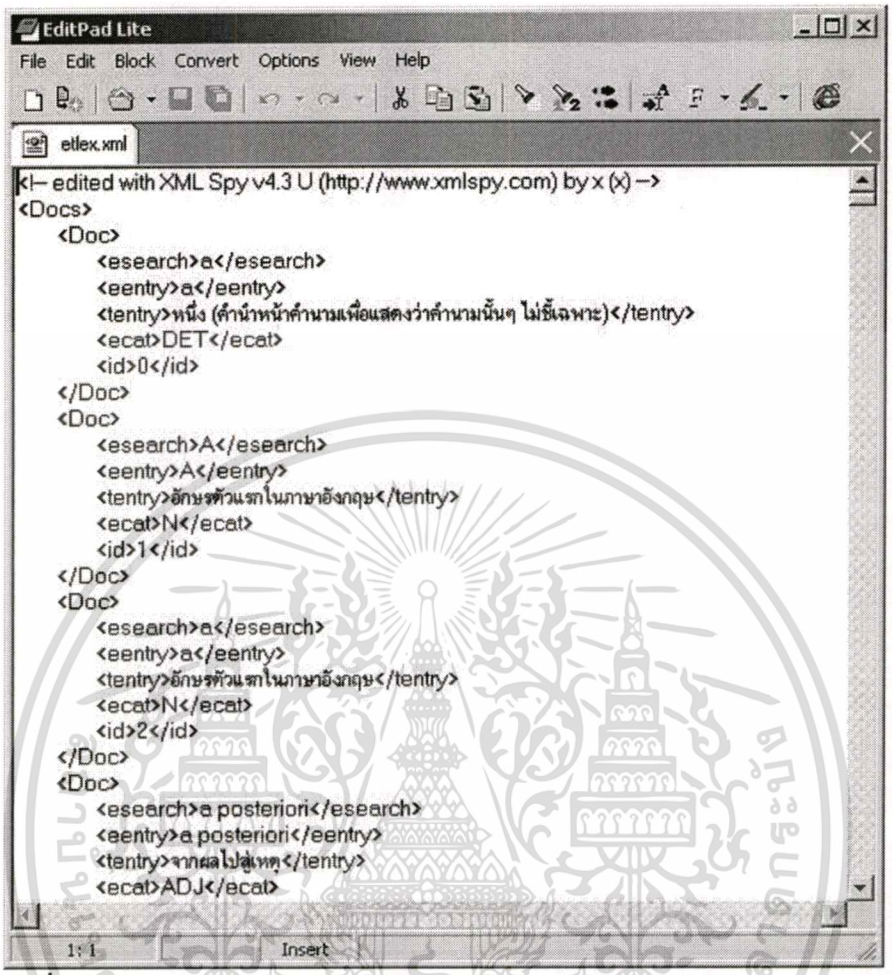
ตัวอย่างบางส่วนของแฟ้มข้อมูล LEXiTRON ก่อนการปรับปรุงได้แสดงไว้ในรูปที่ 3.2 และรูปที่ 3.3 จะแสดงแฟ้มข้อมูลที่ปรับปรุงแล้ว



รูปที่ 3.2 โครงสร้างแฟ้มข้อมูลพจนานุกรม LEXiTRON ก่อนการปรับปรุง

3.1.2 XML (The Extensible Markup Language)

XML เป็นภาษา Markup เชิงข้อความซึ่งเป็นมาตรฐานในการแลกเปลี่ยนข้อมูลบนอินเทอร์เน็ตอย่างรวดเร็ว ผู้ที่ทำหน้าที่รับผิดชอบ และกำหนดมาตรฐานของ XML คือ World Wide Web Consortium (W3C)



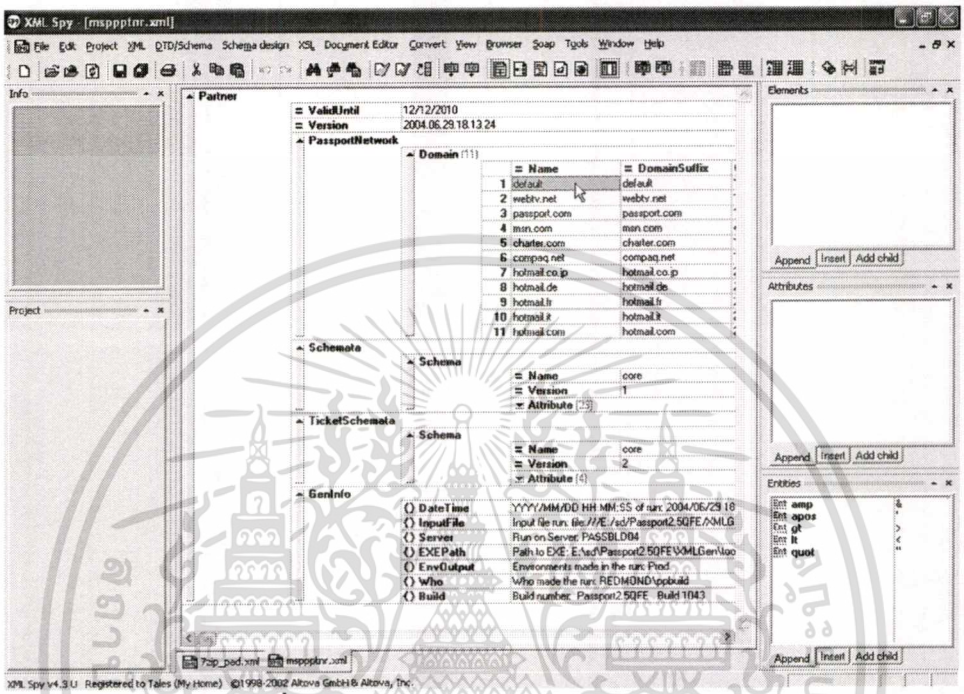
รูปที่ 3.3 โครงสร้างแฟ้มข้อมูลพจนานุกรม LEXiTRON ภายหลังปรับปรุงแล้ว

ฉัตรชัย สุขสอาด (2545) ได้อธิบายความแตกต่างระหว่าง XML และ HTML ไว้ดังนี้ HTML ภาษาที่ใช้ในการเขียนเว็บ มากที่สุดนั่นเป็นเพราะมีรูปแบบที่ง่ายต่อการแสดงผลของ Browser เนื่องจาก มีแท็กตายตัวที่สามารถบอกได้ว่าเมื่อเจอแท็กนี้จะแสดงผลอย่างไร เช่น เมื่อเจอแท็ก <B>...</B> ในเอกสารก็ให้แสดงข้อความที่อยู่ระหว่างแท็กเป็นตัวหนา แต่จะสังเกตได้ว่าคอมพิวเตอร์จะไม่เข้าใจว่าข้อความนั้นคืออะไร เพียงแต่รู้ว่าจะแสดงผลอย่างไร นั่นแสดงว่าไม่สามารถนำข้อมูลภายในแท็กเหล่านี้ไปทำการประมวลใดๆ ได้เลย ในขณะที่ XML เป็นภาษาที่มีลักษะเป็นแท็กคล้าย HTML แต่ไม่ได้มุ่งที่การแสดงผล XML มุ่งที่การสื่อความหมายโดยอนุญาตให้ผู้ใช้สามารถกำหนดแท็กขึ้นได้เองเพื่อให้สื่อความหมายทางภาษาของมนุษย์ แต่คอมพิวเตอร์เองก็เข้าใจเช่นกัน ทำให้ข้อมูลระหว่างแท็กสามารถนำไปประมวลผลต่อได้

3.2 โปรแกรม XML Spy V4.3

XML Spy เป็นโปรแกรมที่ใช้สำหรับการจัดการเอกสาร XML โดยเฉพาะ ซึ่งได้รับความนิยมสูงและมีประสิทธิภาพดี ผู้พัฒนาระบบจึงใช้โปรแกรม XML Spy เพื่อปรับปรุงข้อมูลใน

แฟ้มข้อมูลของ LEXiTRON ให้อยู่ในรูปแบบที่ถูกต้องที่เหมาะสมก่อนที่จะโอนถ่ายข้อมูลไปยังฐานข้อมูลที่ได้จัดเตรียมไว้ ตัวอย่างหน้าจอโปรแกรม XML Spy V4.3 แสดงไว้ดังรูปที่ 3.4



รูปที่ 3.4 หน้าจอโปรแกรม XML Spy V4.3

3.3 Microsoft Visual Studio .NET 2005 BETA2

ผู้พัฒนาต้องการใช้ความสามารถที่เพิ่มขึ้นใน Microsoft .NET Framework 2.0 (เวอร์ชั่นล่าสุด) โดยมีคุณสมบัติเด่นคือช่วยลดการเขียนโปรแกรมลงได้ในการเรียกใช้ฟังก์ชั่นพื้นฐานเมื่อเทียบกับ Visual Studio .NET 2003 มากถึงกว่า 50% (MSDN. 2005) ส่วนภาษาที่จะใช้ในการพัฒนาระบบ ผู้พัฒนาได้เลือกใช้ภาษา ASP.NET และ Visual Basic 2005 ซึ่งยังเป็นเวอร์ชั่น BETA2 ดังแสดงตัวอย่างในรูปที่ 3.5

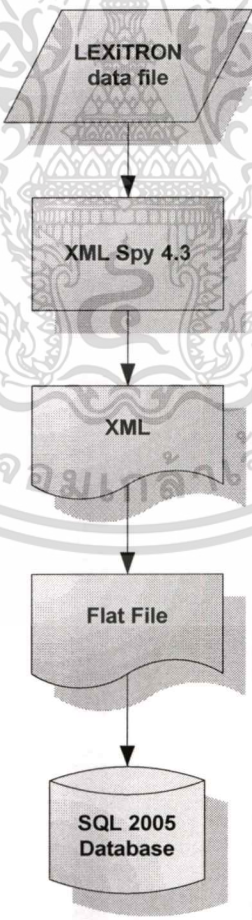
ตารางที่ 3.1 ความสามารถของ Visual Basic .NET 2005 ในการลดการเขียนโปรแกรม

Visual Basic .NET 2003	Const GreetingName As String = "Greeting" Dim sDisplay As Object Dim ResMgr As ResourceManager ResMgr = New ResourceManager("ResourcesSample.MyStrings",_ Me.GetType.Assembly) sDisplay = ResMgr.GetString(GreetingName)
Visual Basic 2005	My.Resources.MyStrings.Greeting

3.4 Microsoft SQL Server 2005 Developer Edition BETA2

ผู้พัฒนาได้เลือกใช้ระบบจัดการฐานข้อมูลนี้เนื่องด้วยในการให้บริการจริงในเครือข่ายอินเทอร์เน็ตจำเป็นต้องเลือกฐานข้อมูลที่รองรับการใช้บริการจำนวนมากได้ ผู้พัฒนาได้คัดเลือกระบบจัดการข้อมูล 2 ชนิด ได้แก่ Microsoft Access 2003 และ Microsoft SQL Server 2005 Developer Edition BETA2 โดยการทดสอบได้เน้นหนักในเรื่องของการอ่านข้อมูล เนื่องจากระบบพจนานุกรม ส่วนใหญ่จะทำการอ่านค่าอย่างเดียว ผลปรากฏว่า Microsoft Access 2003 เร็วกว่าเล็กน้อยในการค้นหาข้อมูลคำศัพท์ในพจนานุกรมที่มีจำนวนคำศัพท์ทั้งภาษาไทยและภาษาอังกฤษจำนวนประมาณ 120,000 คำ แต่เมื่อคำนึงถึงเสถียรภาพในการใช้งานจริงแล้วจึงตัดสินใจเลือกใช้ Microsoft SQL Server 2005 ซึ่งสามารถรองรับการใช้งานอย่างหนักได้ดีกว่า

สำหรับขั้นตอนการนำแฟ้มข้อมูลคำศัพท์จากโปรแกรม LEXiTRON เข้าจัดเก็บในฐานข้อมูล SQL Server 2005 นั้น แสดงไว้ดังรูปที่ 3.5

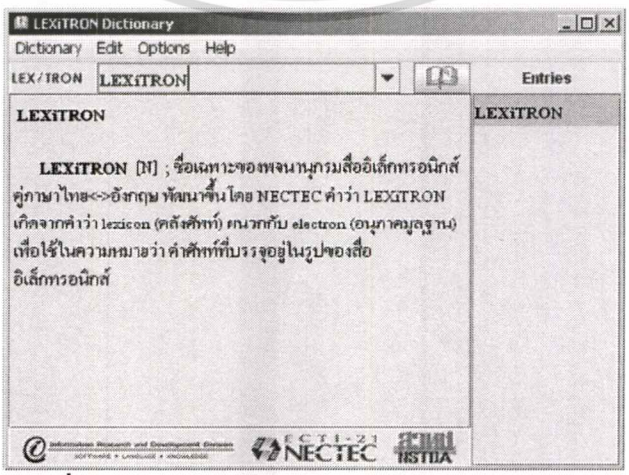


รูปที่ 3.5 ขั้นตอนการนำแฟ้มข้อมูลคำศัพท์จากโปรแกรม LEXiTRON เข้าจัดเก็บในฐานข้อมูล

บทที่ 3
เครื่องมือที่ใช้พัฒนาโครงการ

3.1 พจนานุกรม LEXiTRON

พจนานุกรม LEXiTRON เป็นพจนานุกรมสื่ออิเล็กทรอนิกส์ ไทย - อังกฤษ พัฒนาขึ้นโดยฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ หลักการที่สำคัญของการพัฒนา พจนานุกรมนี้คือ ความต้องการที่จะนำเทคโนโลยีฐานข้อมูลขนาดใหญ่เข้ามาช่วยในการวิจัยและพัฒนาในสาขาการประมวลผลภาษาธรรมชาติ ด้วยแนวคิดที่ว่าฐานข้อมูลขนาดใหญ่สามารถสะท้อนรูปแบบของภาษาโดยรวมได้ เราเรียกการพัฒนาฐานข้อมูลพจนานุกรมในลักษณะนี้ว่า การสร้างพจนานุกรมจากฐานข้อมูลขนาดใหญ่ (Corpus-based dictionary) ซึ่งหลักการของการสร้างฐานข้อมูลชนิดนี้คือ การรวบรวมคำศัพท์จากคำที่ปรากฏใช้จริงในอัตราความถี่สูงในบริบทต่างๆ ของการใช้ภาษา โดยใช้เทคโนโลยีทางคอมพิวเตอร์รวบรวมและคัดเลือกมาจากฐานข้อมูลขนาดใหญ่ จากแนวคิดดังกล่าว ทำให้ LEXiTRON เป็นพจนานุกรมค้นหาศัพท์คู่ภาษา ไทย อังกฤษ ที่ประกอบด้วยศัพท์ทันสมัยที่มีใช้อยู่ในปัจจุบันจำนวนมาก และนอกเหนือจากคุณประโยชน์ที่ได้จากการใช้เทคโนโลยีทางด้านคอมพิวเตอร์ผสมผสานกับภาษาศาสตร์เพื่อประโยชน์ทางการใช้ที่ง่ายและรวดเร็วขึ้นนั้น ยังถือเป็นประโยชน์อย่างยิ่งสำหรับการค้นคว้าข้อมูลเกี่ยวกับคำศัพท์เพื่อใช้ประโยชน์ทางด้านการวิจัยทางภาษาอย่างยิ่ง (ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 2546) ตัวอย่างหน้าจอของโปรแกรมพจนานุกรม LEXiTRON แสดงไว้ดังรูปที่ 3.1



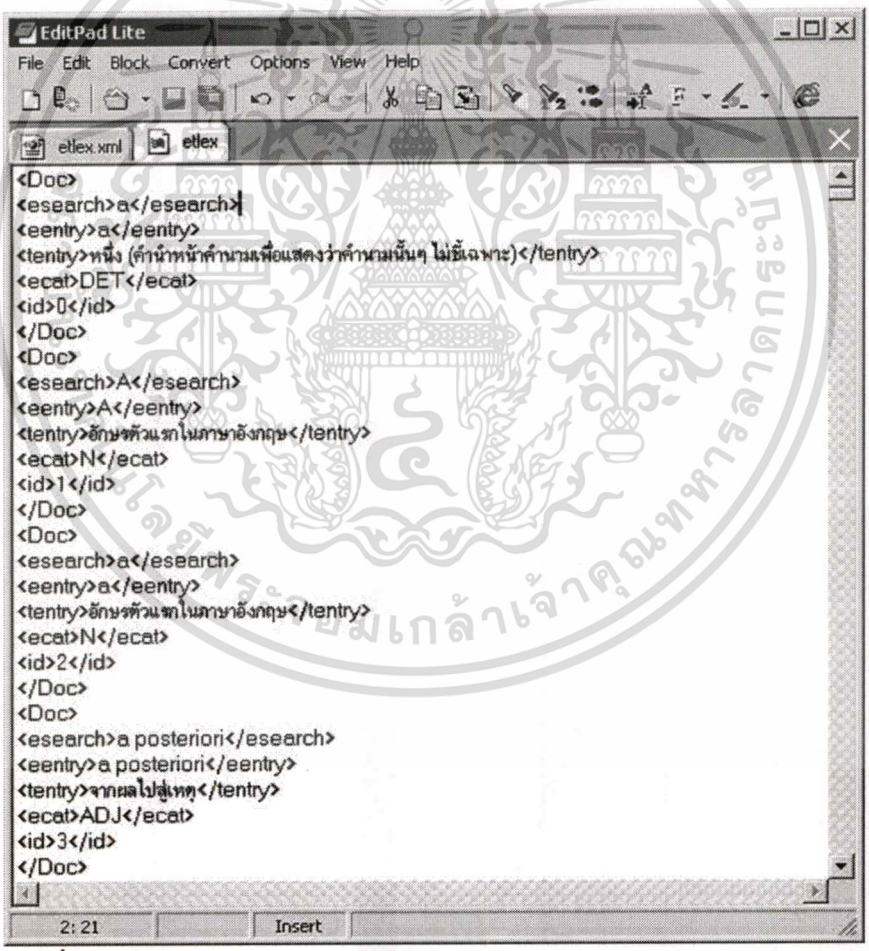
รูปที่ 3.1 หน้าจอโปรแกรมพจนานุกรม LEXiTRON

3.1.1 โครงสร้างของแฟ้มข้อมูล LEXiTRON

โครงสร้างแฟ้มข้อมูลของ LEXiTRON มีลักษณะคล้ายกับเอกสาร XML แต่ยังมีรูปแบบไม่ถูกต้อง ดังนั้นจึงจำเป็นต้องปรับปรุงโครงสร้างและข้อมูลให้อยู่ในรูปแบบของ XML ที่เหมาะสมเสียก่อน จึงจะสามารถนำไปใช้ประโยชน์ต่อไปได้ โดยการปรับปรุงมีสองขั้นตอนได้แก่

1. เพิ่มแท็ก <Docs> … </Docs> ซึ่งเป็นรูทอิลิเมนต์ให้แก่เอกสาร XML
2. การตัดอักขระที่สงวนไว้ออกจากเอกสาร XML

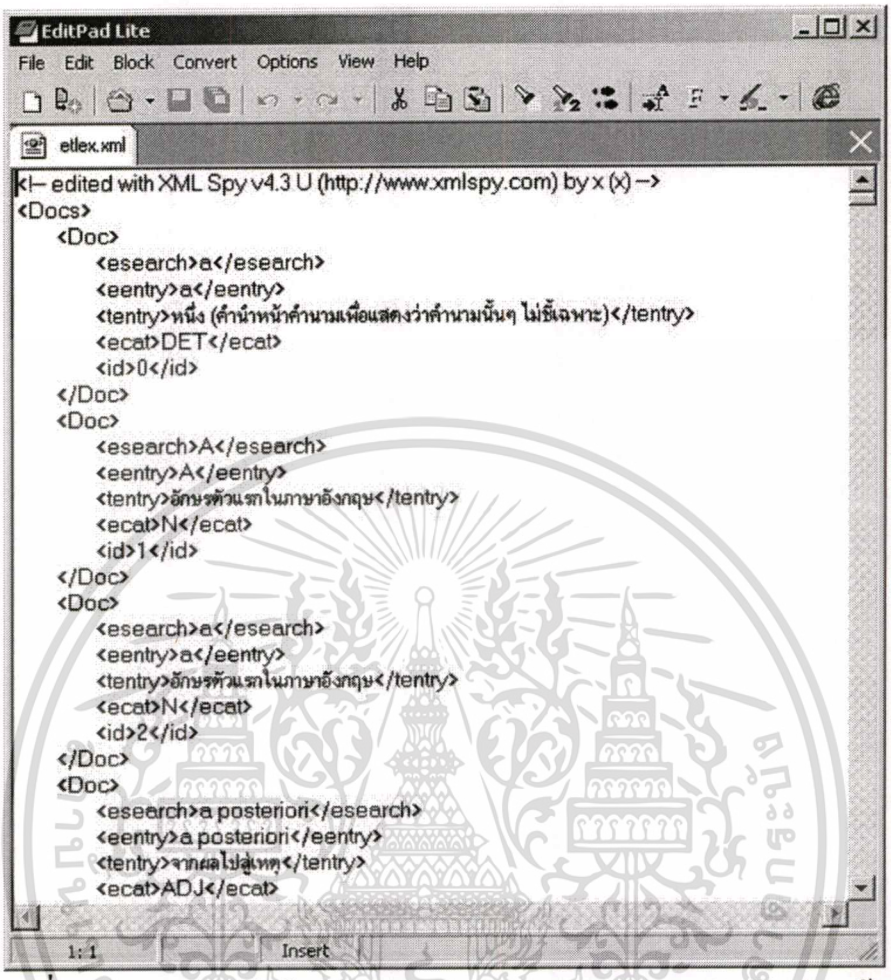
ตัวอย่างบางส่วนของแฟ้มข้อมูล LEXiTRON ก่อนการปรับปรุงได้แสดงไว้ในรูปที่ 3.2 และรูปที่ 3.3 จะแสดงแฟ้มข้อมูลที่ปรับปรุงแล้ว



รูปที่ 3.2 โครงสร้างแฟ้มข้อมูลพจนานุกรม LEXiTRON ก่อนการปรับปรุง

3.1.2 XML (The Extensible Markup Language)

XML เป็นภาษา Markup เชิงข้อความซึ่งเป็นมาตรฐานในการแลกเปลี่ยนข้อมูลบนอินเทอร์เน็ตอย่างรวดเร็ว ผู้ที่ทำหน้าที่รับผิดชอบ และกำหนดมาตรฐานของ XML คือ World Wide Web Consortium (W3C)



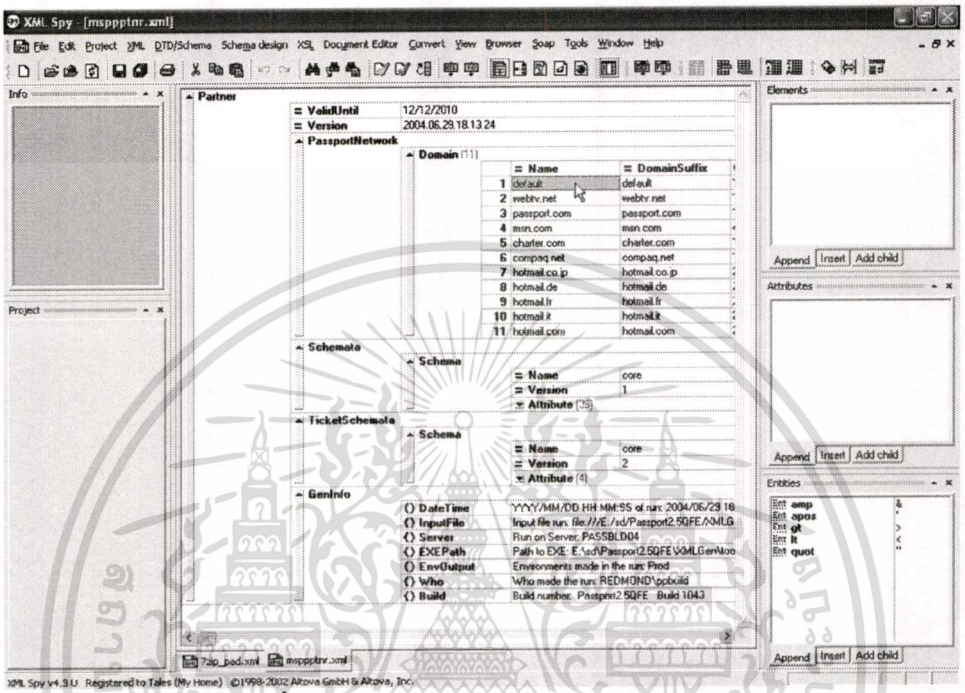
รูปที่ 3.3 โครงสร้างแฟ้มข้อมูลพจนานุกรม LEXiTRON ภายหลังปรับปรุงแล้ว

ฉัตรชัย สุขสอาด (2545) ได้อธิบายความแตกต่างระหว่าง XML และ HTML ไว้ดังนี้ HTML ภาษาที่ใช้ในการเขียนเว็บ มากที่สุดนั่นเป็นเพราะมีรูปแบบที่ง่ายต่อการแสดงผลของ Browser เนื่องจาก มีแท็กตายตัวที่สามารถบอกได้ว่าเมื่อเจอแท็กนี้จะแสดงผลอย่างไร เช่น เมื่อเจอแท็ก <B>...</B> ในเอกสารก็ให้แสดงข้อความที่อยู่ระหว่างแท็กเป็นตัวหนา แต่จะสังเกตได้ว่าคอมพิวเตอร์จะไม่เข้าใจว่าข้อความนั้นคืออะไร เพียงแต่รู้ว่าจะแสดงผลอย่างไร นั่นแสดงว่าไม่สามารถนำข้อมูลภายในแท็กเหล่านี้ไปทำการประมวลใดๆ ได้เลย ในขณะที่ XML เป็นภาษาที่มีลักษะเป็นแท็กคล้าย HTML แต่ไม่ได้มุ่งที่การแสดงผล XML มุ่งที่การสื่อความหมายโดยอนุญาตให้ผู้ใช้สามารถกำหนดแท็กขึ้นได้เองเพื่อให้สื่อความหมายทางภาษาของมนุษย์ แต่คอมพิวเตอร์เองก็เข้าใจเช่นกัน ทำให้ข้อมูลระหว่างแท็กสามารถนำไปประมวลผลต่อได้

3.2 โปรแกรม XML Spy V4.3

XML Spy เป็นโปรแกรมที่ใช้สำหรับการจัดการเอกสาร XML โดยเฉพาะ ซึ่งได้รับความนิยมสูงและมีประสิทธิภาพดี ผู้พัฒนาระบบจึงใช้โปรแกรม XML Spy เพื่อปรับปรุงข้อมูลใน

แฟ้มข้อมูลของ LEXiTRON ให้อยู่ในรูปแบบที่ถูกต้องที่เหมาะสมก่อนที่จะโอนถ่ายข้อมูลไปยังฐานข้อมูลที่ได้จัดเตรียมไว้ ตัวอย่างหน้าจอโปรแกรม XML Spy V4.3 แสดงไว้ดังรูปที่ 3.4



รูปที่ 3.4 หน้าจอโปรแกรม XML Spy V4.3

3.3 Microsoft Visual Studio .NET 2005 BETA2

ผู้พัฒนาต้องการใช้ความสามารถที่เพิ่มขึ้นใน Microsoft .NET Framework 2.0 (เวอร์ชั่นล่าสุด) โดยมีคุณสมบัติเด่นคือช่วยลดการเขียนโปรแกรมลงได้ในการเรียกใช้ฟังก์ชั่นพื้นฐานเมื่อเทียบกับ Visual Studio .NET 2003 มากถึงกว่า 50% (MSDN. 2005) ส่วนภาษาที่จะใช้ในการพัฒนาระบบ ผู้พัฒนาได้เลือกใช้ภาษา ASP.NET และ Visual Basic 2005 ซึ่งยังเป็นเวอร์ชั่น BETA2 ดังแสดงตัวอย่างในรูปที่ 3.5

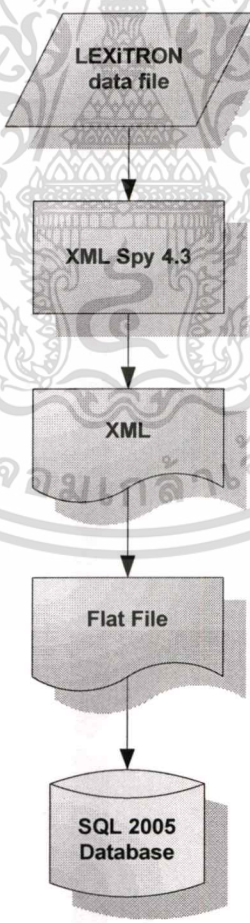
ตารางที่ 3.1 ความสามารถของ Visual Basic .NET 2005 ในการลดการเขียนโปรแกรม

Visual Basic .NET 2003	Const GreetingName As String = "Greeting" Dim sDisplay As Object Dim ResMgr As ResourceManager ResMgr = New ResourceManager("ResourcesSample.MyStrings",_ Me.GetType.Assembly) sDisplay = ResMgr.GetString(GreetingName)
Visual Basic 2005	My.Resources.MyStrings.Greeting

3.4 Microsoft SQL Server 2005 Developer Edition BETA2

ผู้พัฒนาได้เลือกใช้ระบบจัดการฐานข้อมูลนี้เนื่องด้วยในการให้บริการจริงในเครือข่ายอินเทอร์เน็ตจำเป็นต้องเลือกฐานข้อมูลที่รองรับการใช้บริการจำนวนมากได้ ผู้พัฒนาได้คัดเลือกระบบจัดการข้อมูล 2 ชนิด ได้แก่ Microsoft Access 2003 และ Microsoft SQL Server 2005 Developer Edition BETA2 โดยการทดสอบได้เน้นหนักในเรื่องของการอ่านข้อมูล เนื่องจากระบบพจนานุกรม ส่วนใหญ่จะทำการอ่านค่าอย่างเดียว ผลปรากฏว่า Microsoft Access 2003 เร็วกว่าเล็กน้อยในการค้นหาข้อมูลคำศัพท์ในพจนานุกรมที่มีจำนวนคำศัพท์ทั้งภาษาไทยและภาษาอังกฤษจำนวนประมาณ 120,000 คำ แต่เมื่อคำนึงถึงเสถียรภาพในการใช้งานจริงแล้วจึงตัดสินใจเลือกใช้ Microsoft SQL Server 2005 ซึ่งสามารถรองรับการใช้งานอย่างหนักได้ดีกว่า

สำหรับขั้นตอนการนำแฟ้มข้อมูลคำศัพท์จากโปรแกรม LEXiTRON เข้าจัดเก็บในฐานข้อมูล SQL Server 2005 นั้น แสดงไว้ดังรูปที่ 3.5



รูปที่ 3.5 ขั้นตอนการนำแฟ้มข้อมูลคำศัพท์จากโปรแกรม LEXiTRON เข้าจัดเก็บในฐานข้อมูล

บทที่ 4

การวิเคราะห์และออกแบบระบบ

4.1 การวิเคราะห์ระบบงาน

จากการศึกษาทฤษฎีต่างๆ ที่เกี่ยวข้อง และเพื่อให้บรรลุตามวัตถุประสงค์ของการพัฒนาระบบ ผู้พัฒนาระบบสามารถแบ่งการวิเคราะห์ระบบออกเป็น 2 ขั้นตอน ได้แก่

1. การวิเคราะห์ขั้นตอนการทำงานของระบบ
2. การวิเคราะห์ขั้นตอนการจัดเตรียมฐานข้อมูลพจนานุกรม

4.1.1 การวิเคราะห์ขั้นตอนการทำงานของระบบ

1. การทำงานเริ่มจากการรับคำศัพท์ที่ใช้เป็นข้อคำถาม เลือกภาษาของข้อคำถาม เลือกภาษาของคำศัพท์ที่ต้องการเป็นผลลัพธ์ และกำหนดค่าอัตราส่วนที่ยอมรับได้โดยคิดเป็นเปอร์เซนต์ เมื่อเปรียบเทียบกับคำที่เป็นข้อคำถามหรืออินพุต
2. เลือกวิธีการค้นหา โดยแบ่งเป็น ค้นหาจากเสียง (Sound-like) หรือ ค้นหาจากความหมาย (Meaning)
3. ในกรณีที่เป็นการค้นหาจากเสียง ระบบจะนำข้อคำถามมาทำการเข้ารหัส โดยการเข้ารหัสมีอยู่ 2 ลักษณะ ดังนี้
   - กรณีของคำไทยจะผ่านกระบวนการแปลงรูปคำไทยโดยอาศัยหลักการแปลงเสียง (Romanization Rules Process)
   - กรณีของคำภาษาอังกฤษจะผ่านกระบวนการแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทย โดยอาศัยหลักการแปลงอักษร (Transliteration Rules Process)

   ในกรณีที่เป็นการค้นหาตามความหมายระบบจะค้นหาจากความหมายในฐานข้อมูลพจนานุกรม โดยจะข้ามขั้นตอนที่ 4 และ 5 จากนั้นจะแสดงผลลัพธ์เป็นรายการคำศัพท์ที่มีความหมายตรงกับข้อคำถาม
4. นำข้อคำที่เข้ารหัสแล้วมาผ่านกระบวนการเปรียบเทียบและคัดเลือก โดยการคำนวนหาค่าความคล้ายคลึง (Similarity) ของรหัสคำของคำข้อคำถามกับรหัสคำในพจนานุกรม แล้วเปรียบเทียบกับค่าอัตราส่วนที่ยอมรับได้โดยคิดเป็นเปอร์เซนต์
5. ผลลัพธ์ซึ่งได้แก่คำศัพท์ที่ผ่านเกณฑ์การเปรียบเทียบจะถูกนำมารวบรวม และแสดงผล

6. ผู้ใช้เลือกคำจากรายการผลลัพธ์ที่ได้จากขั้นตอนก่อนหน้า เพื่อแปลความหมายจากฐานข้อมูลพจนานุกรม และแสดงข้อมูลของคำศัพท์ที่ถูกเลือก เช่น แสดงคำแปล หน้าที่ของคำศัพท์ เป็นต้น

4.1.2 การวิเคราะห์ขั้นตอนการจัดเตรียมฐานข้อมูลพจนานุกรม

หลังจากที่ได้วิเคราะห์ขั้นตอนต่าง ๆ ที่จะเกิดขึ้นในการจัดเตรียมฐานข้อมูลพจนานุกรม ผู้พัฒนาสามารถแบ่งการทำงานออกเป็น 2 ขั้นตอนดังนี้

1. นำข้อมูลจากแฟ้มข้อมูลพจนานุกรม LEXiTRON เข้ามาเก็บในฐานข้อมูลที่จัดเตรียมไว้
2. เข้ารหัสคำศัพท์แล้วจัดเก็บกลับไปยังฐานข้อมูลพจนานุกรม โดยการเข้ารหัสมีอยู่ 2 ลักษณะ ดังนี้
   - กรณีของคำไทยจะผ่านกระบวนการแปลงรูปคำไทยโดยอาศัยหลักการแปลงเสียง (Romanization Rules Process)
   - กรณีของคำภาษาอังกฤษจะผ่านกระบวนการแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทย โดยอาศัยหลักการแปลงอักษร (Transliteration Rules Process)

4.2 การออกแบบระบบงาน

หลังจากทำการวิเคราะห์ระบบแล้ว ผู้พัฒนาสามารถแบ่งกระบวนการออกแบบระบบออกเป็น 4 ส่วน ดังนี้

1. การออกแบบระบบการเข้ารหัสคำ เปรียบเทียบรหัสเสียง และค้นคืนข้อมูลพจนานุกรมด้วยคำทับศัพท์หรือคำพ้องเสียง
2. การออกแบบฐานข้อมูลพจนานุกรม
3. การออกแบบสถาปัตยกรรมเครือข่ายของระบบงาน
4. การออกแบบสถาปัตยกรรมของระบบเครือข่าย

4.2.1 การออกแบบระบบการเข้ารหัสคำ เปรียบเทียบรหัสเสียง และค้นคืนข้อมูลพจนานุกรมด้วยคำทับศัพท์หรือคำพ้องเสียง

ผู้พัฒนาสามารถสรุปหน้าที่หรือฟังก์ชั่นการทำงานของระบบได้ 5 ส่วน ดังต่อไปนี้

1. ฟังก์ชั่นการเข้ารหัสคำไทยทับศัพท์คำอังกฤษให้เป็นรหัสแทนเสียง
2. ฟังก์ชั่นการเข้ารหัสคำอังกฤษทับศัพท์คำไทยให้เป็นรหัสแทนเสียง
3. ฟังก์ชั่นการเข้ารหัสคำศัพท์ในฐานข้อมูลพจนานุกรม
4. ฟังก์ชั่นการเปรียบเทียบคำแบบประมาณ
5. ฟังก์ชั่นการค้นหาและแสดงผลความหมายของคำศัพท์

4.2.1.1 การออกแบบฟังก์ชั่นการเข้ารหัสคำไทยทับศัพท์คำอังกฤษให้เป็นรหัสแทนเสียง

ขั้นตอนการเข้ารหัสคำไทยที่ทับศัพท์คำอังกฤษหรือการแปลงรูปคำไทย มีจุดประสงค์หลักเพื่อแปลงคำไทยที่อ่านออกเสียงคล้ายกัน แต่เขียนได้หลายรูปแบบ ให้อยู่ในรูปแบบเดียวกัน เพื่อให้ขั้นตอนการเทียบรหัสกระทำได้ง่ายขึ้น การแปลงรูปประกอบด้วยการตัดวรรณยุกต์และไม้ไต่คู้ การเปลี่ยนรูปการสะกด และการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล โดยแสดงขั้นตอนการทำงานดังรูปที่ 4.1

รูปที่ 4.1 ขั้นตอนการทำงานของฟังก์ชั่นการเข้ารหัสคำไทยทับศัพท์คำอังกฤษให้เป็นรหัสแทนเสียง

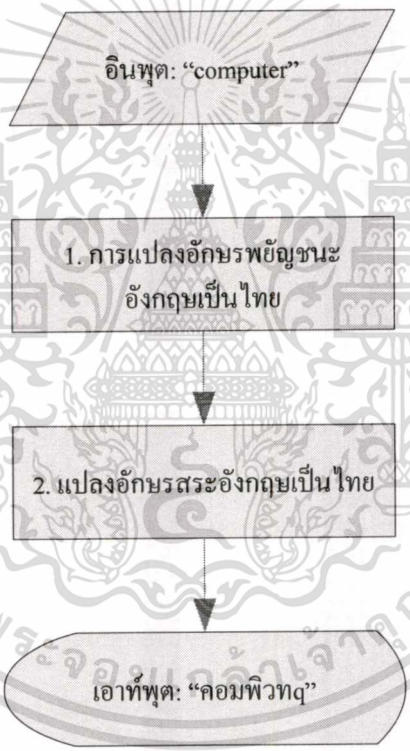
4.2.1.2 การออกแบบฟังก์ชั่นการเข้ารหัสคำอังกฤษทับศัพท์คำไทยให้เป็นรหัสแทนเสียง

ในกรณีที่ข้อคำถามเป็นคำอังกฤษทับศัพท์คำไทย จะทำการแปลงอักษรอังกฤษเป็นไทย การแปลงอักษรที่นำเสนอในที่นี้เป็นการเปลี่ยนพยัญชนะอังกฤษเป็นไทย ส่วนสระอังกฤษจะใช้หลักเกณฑ์การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล คือจะแปลงสระอังกฤษเป็นสระไทย แต่ถ้าสระไทยนั้นเป็นสระที่ใช้อักขระตั้งแต่สองตัวขึ้นไป จะใช้สัญลักษณ์เสียงสากล แทนเสียงสระ

ดังกล่าว หลักเกณฑ์ในการแปลงอักษรในส่วนพยัญชนะ (ตามตารางที่ 2.1) โดยแสดงขั้นตอนการทำงานดังรูปที่ 4.2

4.2.1.3 การออกแบบฟังก์ชั่นการเข้ารหัสคำศัพท์ในฐานข้อมูลพจนานุกรม

หลังจากที่ได้นำข้อมูลจากแฟ้มข้อมูลพจนานุกรม LEXiTRON เข้ามาเก็บในฐานข้อมูล SQL Server ที่จัดเตรียมไว้ จะนำข้อมูลในฐานข้อมูลมาทำการเข้ารหัสคำในพจนานุกรมทั้งคำไทยและอังกฤษ ให้เป็นรหัสแทนเสียงของแต่ละคำ แล้วจัดเก็บเข้าไปยังฐานข้อมูลพจนานุกรมอีกครั้ง โดยแสดงขั้นตอนการทำงานดังรูปที่ 4.3



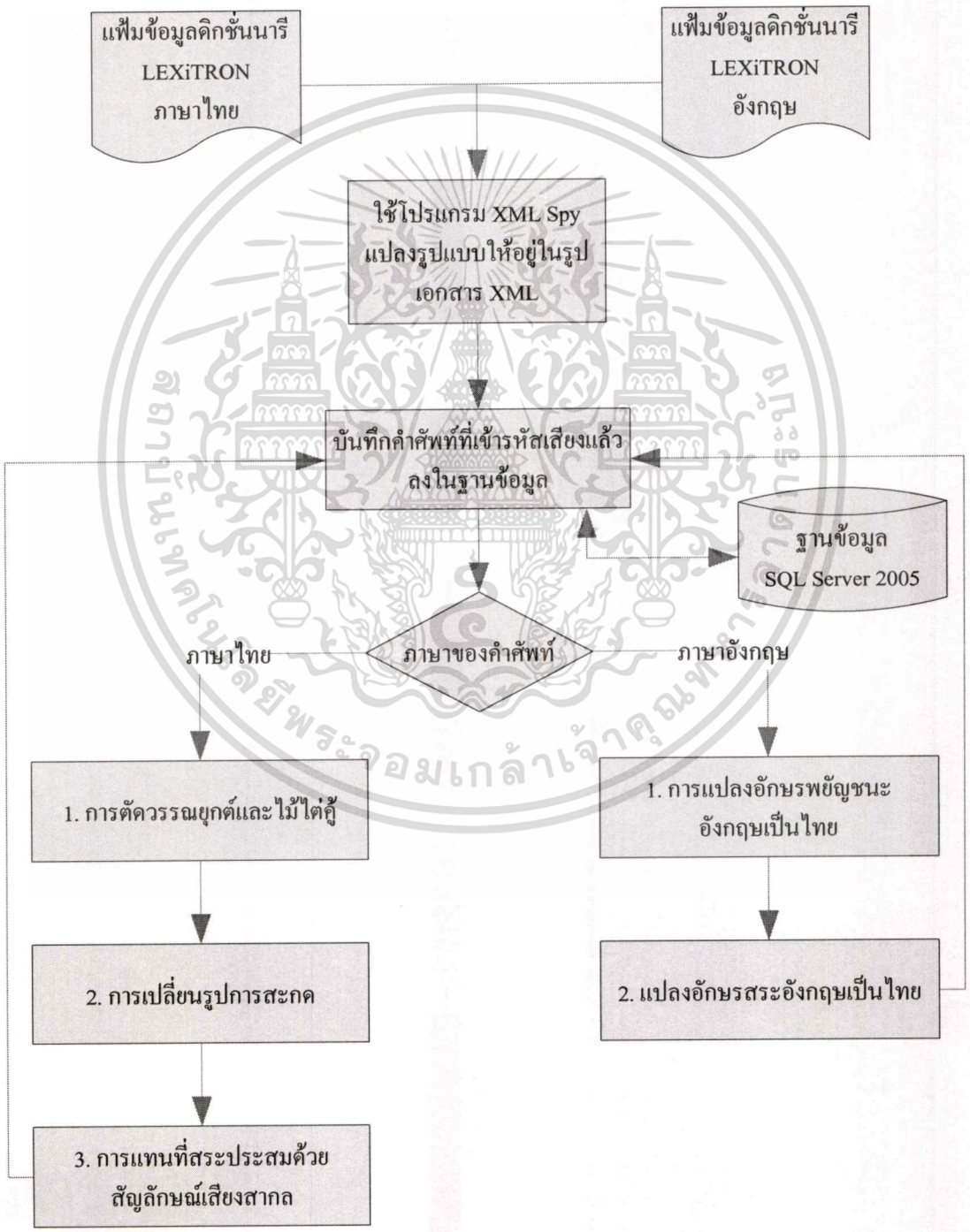
รูปที่ 4.2 ขั้นตอนการทำงานของฟังก์ชั่นการเข้ารหัสคำอังกฤษทับศัพท์คำไทยให้เป็นรหัสแทนเสียง

4.2.1.4 การออกแบบฟังก์ชั่นการเปรียบเทียบคำแบบประมาณ

รหัสคำที่ได้ของคู่คำไทยและคำภาษาอังกฤษทับศัพท์คำไทยนั้น อาจไม่ตรงกันทุกตัวอักษร แต่จะมีลักษณะคล้ายกัน การเทียบรหัสคำแบบประมาณโดยอาศัยการคำนวณค่าความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคที่เรียกว่า N-Gram Based Techniques จากนั้น นำค่าความแตกต่างของคู่รหัสคำที่ได้มาทดสอบกับเงื่อนไขในการเปรียบเทียบ ถ้าผ่านการทดสอบจะสรุปได้ว่ารหัสคำทั้งสองรหัสเป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา โดยแสดงขั้นตอนการทำงานดังรูปที่ 4.4

4.2.1.5 การออกแบบฟังก์ชั่นการค้นหาและแสดงผลความหมายของคำศัพท์

หลังจากที่ผู้ใช้เลือกคำศัพท์ที่ผ่านเกณฑ์การเปรียบเทียบแล้ว ระบบจะค้นหาความหมายของคำศัพท์นั้นจากฐานข้อมูลพจนานุกรม โดยใช้คำศัพท์ที่ผู้ใช้เลือกเป็นคีย์ในการค้นหา โดยแสดงขั้นตอนการทำงานดังรูปที่ 4.5



รูปที่ 4.3 ขั้นตอนการทำงานของฟังก์ชั่นการเข้ารหัสคำศัพท์ในฐานข้อมูลพจนานุกรม

4.2.2 การออกแบบฐานข้อมูลพจนานุกรม

เนื่องจากคำศัพท์สามารถมีได้หลายความหมาย ทำให้ต้องเก็บระเบียนที่อาจมีคำศัพท์ที่ซ้ำกันได้ ผู้พัฒนาระบบจึงเลือกใช้ Non-clustered Index บนคอลัมน์ Word เพื่อช่วยเพิ่มความเร็วในการค้นหาความหมาย สำหรับการปรับปรุงข้อมูลทำได้โดยใช้ภาษา SQL ในการเพิ่มเติม แก้ไข หรือลบ ข้อมูลคำศัพท์ในฐานข้อมูลพจนานุกรม เนื่องจากระบบยังไม่รองรับการทำงานในส่วนนี้ โดยโครงสร้างฐานข้อมูลแสดงไว้ในตารางที่ 4.1 และรูปที่ 4.6

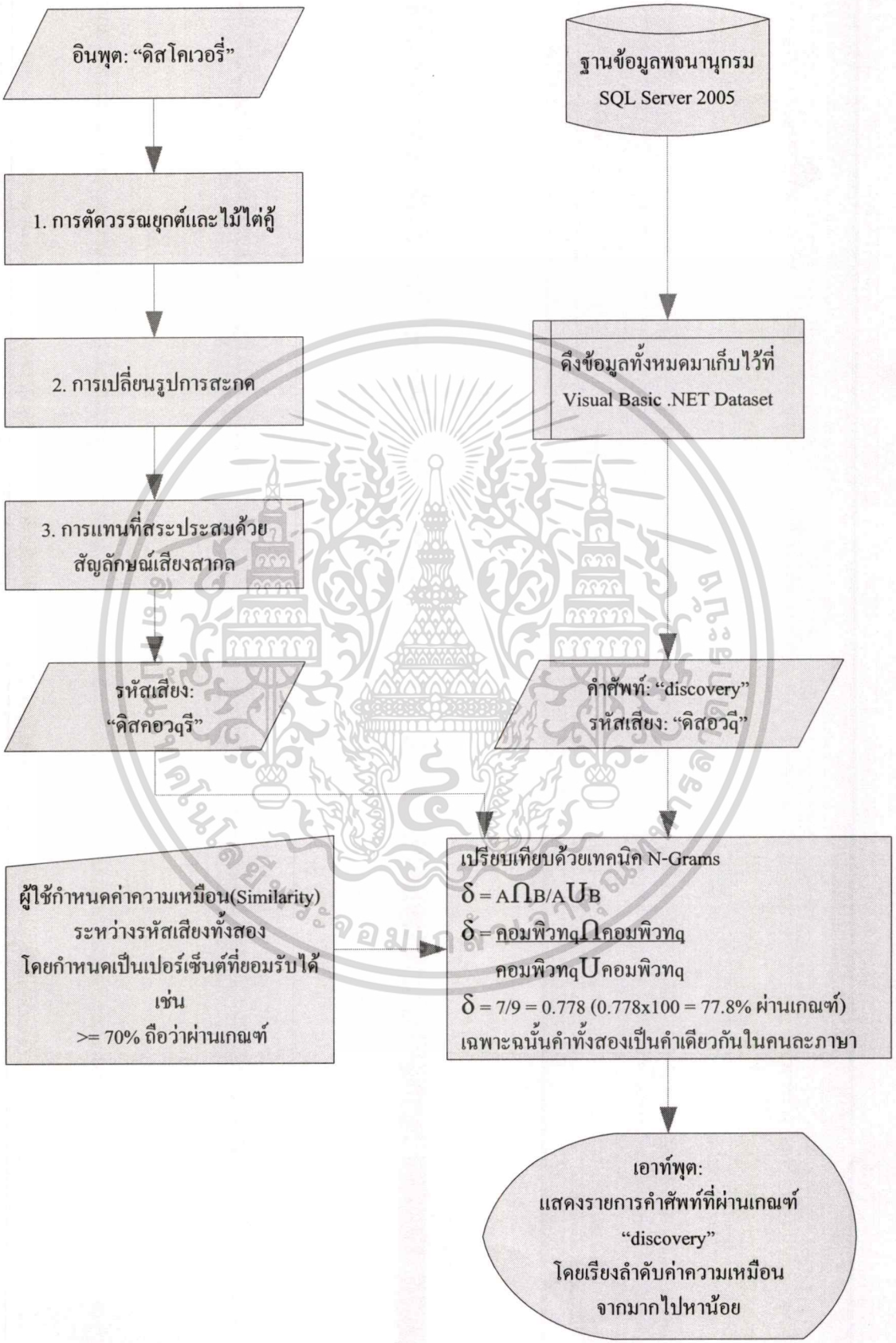
4.2.3 การออกแบบสถาปัตยกรรมของระบบงาน

การออกแบบสถาปัตยกรรมของระบบงานจะเป็นรูปแบบ 3 เทียร์ ซึ่งจะประกอบด้วย 4 องค์ประกอบ ดังแสดงโครงสร้างการทำงานในรูปที่ 4.7

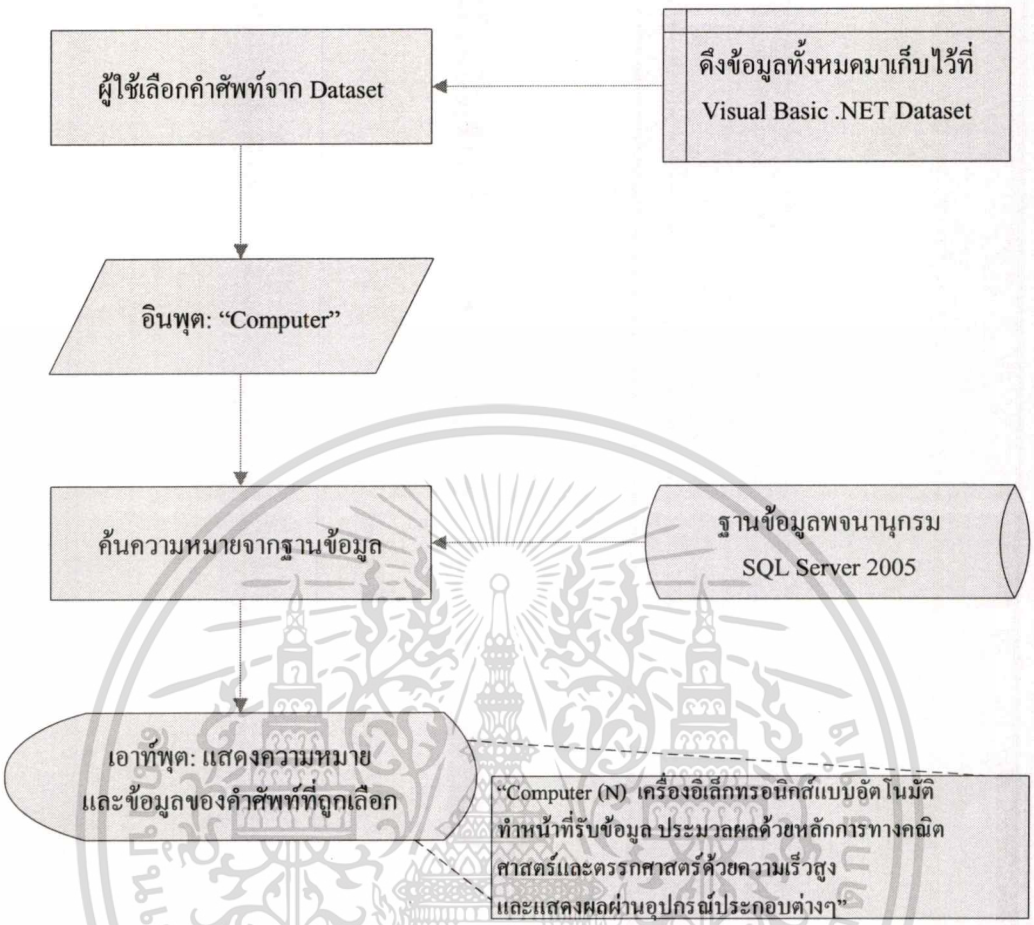
1. **ไคลเอนท์เว็บเบราเซอร์** ใช้ในการแสดงผลและติดต่อระหว่างกับผู้ใช้กับระบบ
2. **เว็บเซิร์ฟเวอร์** ให้บริการข้อมูลแก่ไคลเอนท์เว็บเบราเซอร์ ในรูปแบบของ HTML
3. **แอพพลิเคชั่นเซิร์ฟเวอร์** ใช้ในการประมวลผล ซึ่งได้แก่ การเข้ารหัสเสียง การเปรียบเทียบให้คะแนน และส่งข้อมูลผลลัพธ์จากการทำงานกลับไปยังเว็บเซิร์ฟเวอร์ เพื่อแปลงข้อมูลให้อยู่ในรูปแบบของ HTML เพื่อแสดงผลในไคลเอนท์เว็บเบราเซอร์ ต่อไป
4. **เดต้าเบสเซิร์ฟเวอร์** ใช้ในการจัดเก็บและให้บริการข้อมูลคำศัพท์แก่ แอพพลิเคชั่นเซิร์ฟเวอร์

4.2.4 การออกแบบสถาปัตยกรรมของระบบเครือข่าย

เนื่องการพัฒนาระบบมุ่งเน้นที่จะให้บริการกับผู้ใช้งานผ่านทางเครือข่ายอินเทอร์เน็ต ดังนั้น จำเป็นต้องออกแบบสถาปัตยกรรมของระบบเครือข่ายที่จะใช้ด้วย โดยผู้ใช้จะติดต่อกับระบบผ่านทางไคลเอนท์เว็บเบราเซอร์ที่เชื่อมต่อกับเครือข่ายอินเทอร์เน็ต โดยจะติดต่อเข้ามายังเซิร์ฟเวอร์ที่ให้บริการอยู่ผ่านทางเร้าเตอร์ สำหรับการเชื่อมต่ออุปกรณ์เครือข่ายในระบบได้แสดงไว้ ดังรูปที่ 4.8



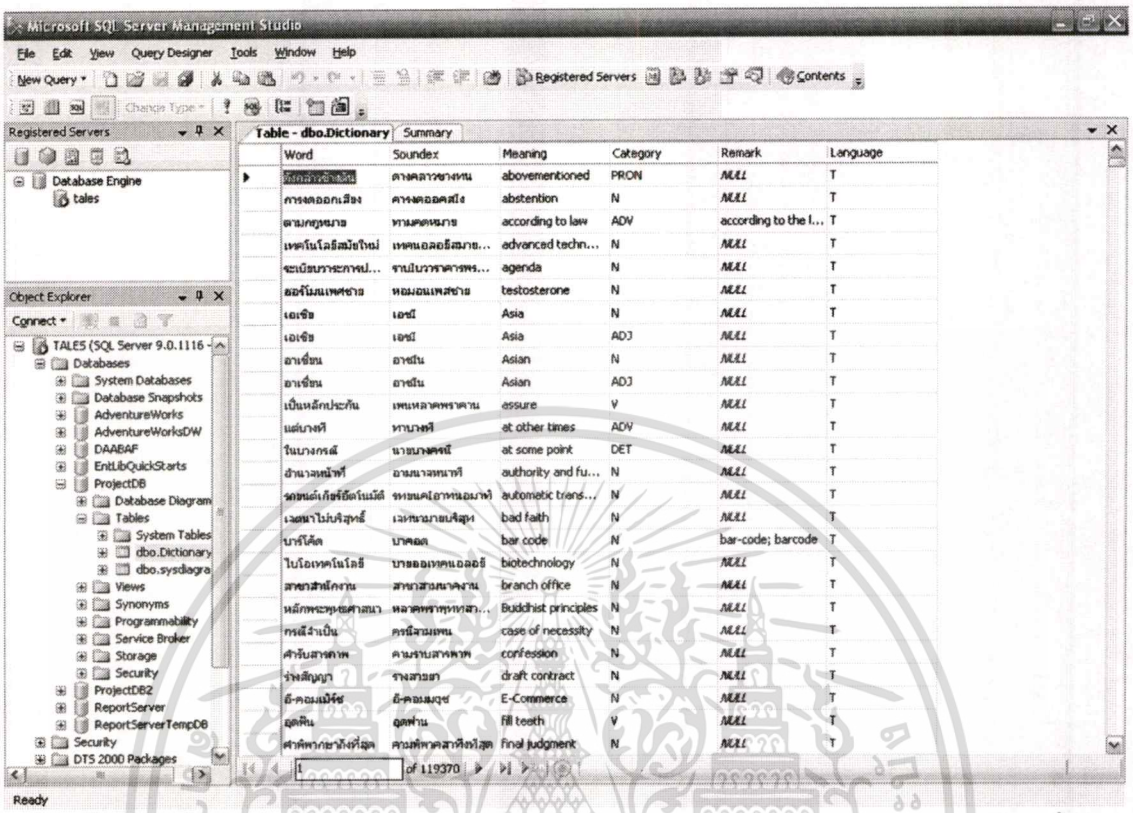
รูปที่ 4.4 ขั้นตอนการทำงานของฟังก์ชั่นการเปรียบเทียบคำแบบประมาณ



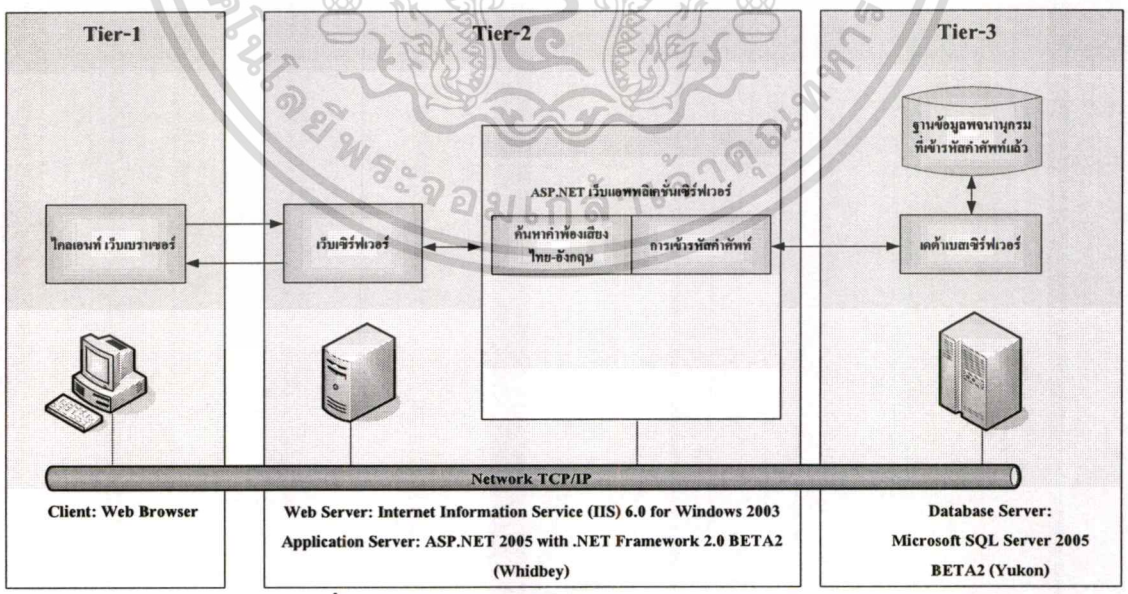
รูปที่ 4.5 ขั้นตอนการทำงานของฟังก์ชั่นการค้นหาและแสดงผลความหมายของคำศัพท์

ตารางที่ 4.1 โครงสร้างตารางที่ออกแบบสำหรับใช้จัดเก็บข้อมูลพจนานุกรม

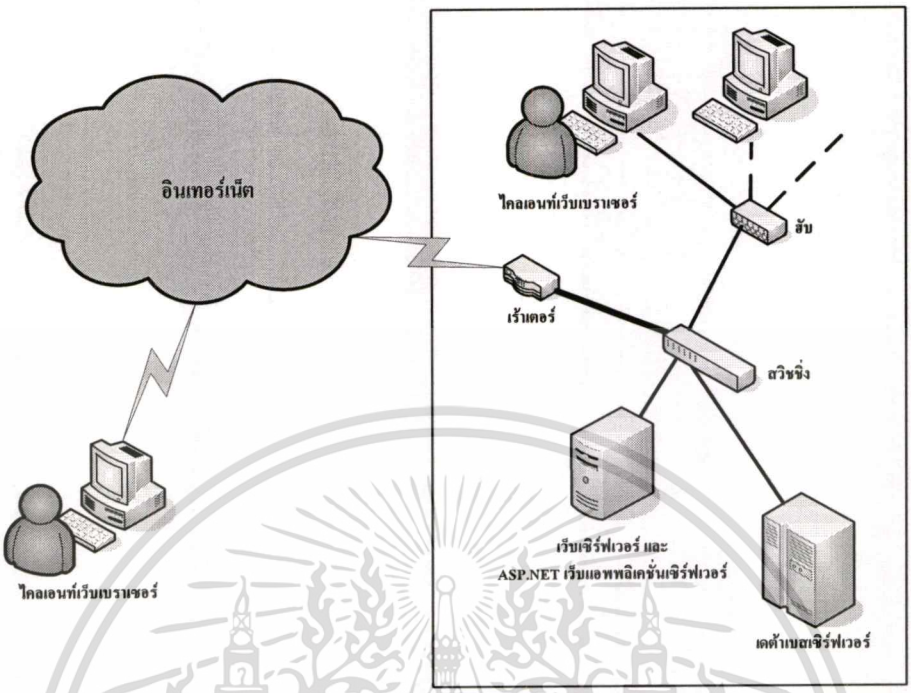
Field Name	Description	Data Type	Allow NULL	Index Type
Word	คำศัพท์	nvarchar(50)	No	Non-clustered Index
Soundex	คำศัพท์ที่เข้ารหัสแล้ว	nvarchar(50)	Yes	
Meaning	ความหมาย	Nvarchar(MAX)	Yes	
Category	ประเภทของคำศัพท์ เช่น คำนาม (N.) คำกริยา (VT.)	nvarchar(10)	Yes	
Remark	คำอธิบายเพิ่มเติม	nvarchar(MAX)	Yes	
Language	ภาษาของคำศัพท์ ไทย = T อังกฤษ = E	nchar(1)	No	



รูปที่ 4.6 เครื่องมือของ Microsoft SQL Server 2005 แสดงโครงสร้างและข้อมูลในตารางชื่อ Dictionary ซึ่งใช้จัดเก็บฐานข้อมูลพจนานุกรม



รูปที่ 4.7 โครงสร้างสถาปัตยกรรมของระบบงาน



รูปที่ 4.8 สถาปัตยกรรมของระบบเครือข่าย

บทที่ 5
การพัฒนาระบบงาน

5.1 การพัฒนาระบบ

ผู้พัฒนาระบบเลือกใช้โปรแกรม Microsoft Visual Studio 2005 BETA2 เพื่อใช้ในการพัฒนาระบบงาน โดยเขียนด้วยภาษา ASP .NET 2005 และใช้ภาษา Visual Basic .NET 2005 ในการเขียนโค้ดเพื่อการทำงานเบื้องหลัง (Code Behind) และใช้โปรแกรม Microsoft SQL Server 2005 Developer Edition BETA2 เป็นระบบจัดการฐานข้อมูลซึ่งใช้สำหรับจัดเก็บฐานข้อมูลพจนานุกรม ในส่วนของระบบปฏิบัติการได้เลือกใช้ Microsoft Windows 2003 Enterprise Server ซึ่งติดตั้ง Internet Information Service เวอร์ชั่น 6.0 ทำหน้าที่เป็นเว็บเซิร์ฟเวอร์ โดยติดตั้ง .NET Framework เวอร์ชั่น 2.0 BETA2

ส่วนของฐานข้อมูลพจนานุกรม ผู้พัฒนาระบบได้นำข้อมูลจากโปรแกรมพจนานุกรม LEXiTRON ที่พัฒนาขึ้นโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ โดยประกอบด้วย คำศัพท์ไทยประมาณสี่หมื่นคำ และคำศัพท์ภาษาอังกฤษประมาณแปดหมื่นคำ

5.1.1 ซอฟต์แวร์ (Software)

ระบบปฏิบัติการ	Microsoft Windows 2003 Server Enterprise Edition with SP1
เครื่องมือพัฒนาโปรแกรม	Microsoft Visual Studio 2005 BETA2
ระบบจัดการฐานข้อมูล	Microsoft SQL Server 2005 Developer Edition BETA2
ฐานข้อมูลพจนานุกรม	LEXiTRON Dictionary version 2.1
ระบบจัดการเอกสาร XML	XML Spy version 4.3

5.1.2 ฮาร์ดแวร์ (Hardware)

ซีพียู	IBM R40 Pentium M 1.3 GHz
แรม	512 MB
ฮาร์ดดิสก์	60 GB

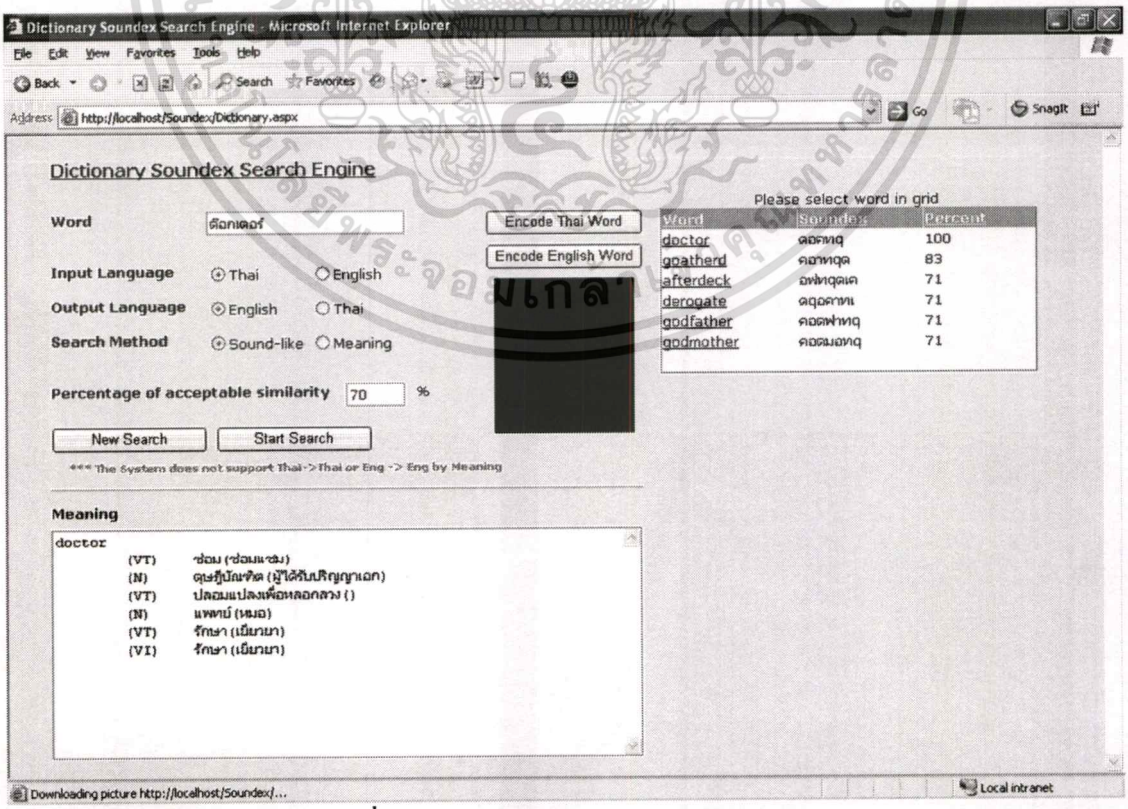
5.2 ฟังก์ชั่นการทำงานของระบบ

ฟังก์ชั่นการทำงานของระบบ สามารถแบ่งได้ 5 ฟังก์ชั่น ดังนี้

1. ฟังก์ชั่นการเข้ารหัสคำศัพท์ภาษาอังกฤษ
2. ฟังก์ชั่นการเข้ารหัสคำศัพท์ภาษาไทย
3. ฟังก์ชั่นการเปรียบเทียบ และวัดค่าความเหมือนของรหัสเสียง
4. ฟังก์ชั่นการค้นหาคำที่ผ่านเกณฑ์ที่กำหนด เพื่อค้นหาคำศัพท์ตามคำที่ออกเสียงคล้ายกัน โดยสามารถค้นหาได้ทั้งหมด 4 รูปแบบ ได้แก่ 1) ไทย-อังกฤษ 2) อังกฤษ-ไทย 3) ไทย-ไทย และ 4) อังกฤษ-อังกฤษ
5. ฟังก์ชั่นการค้นหาตามความหมาย สามารถค้นหาได้ 2 รูปแบบ ได้แก่ 1) ไทย-อังกฤษ และ 2) อังกฤษ-ไทย

5.3 หน้าจอการทำงานหลักของระบบ

หลังจากที่ผู้พัฒนาได้ออกแบบและพัฒนาระบบ ทำให้ได้หน้าจอการทำงานหลักของระบบ ดังแสดงไว้ในรูปที่ 5.1



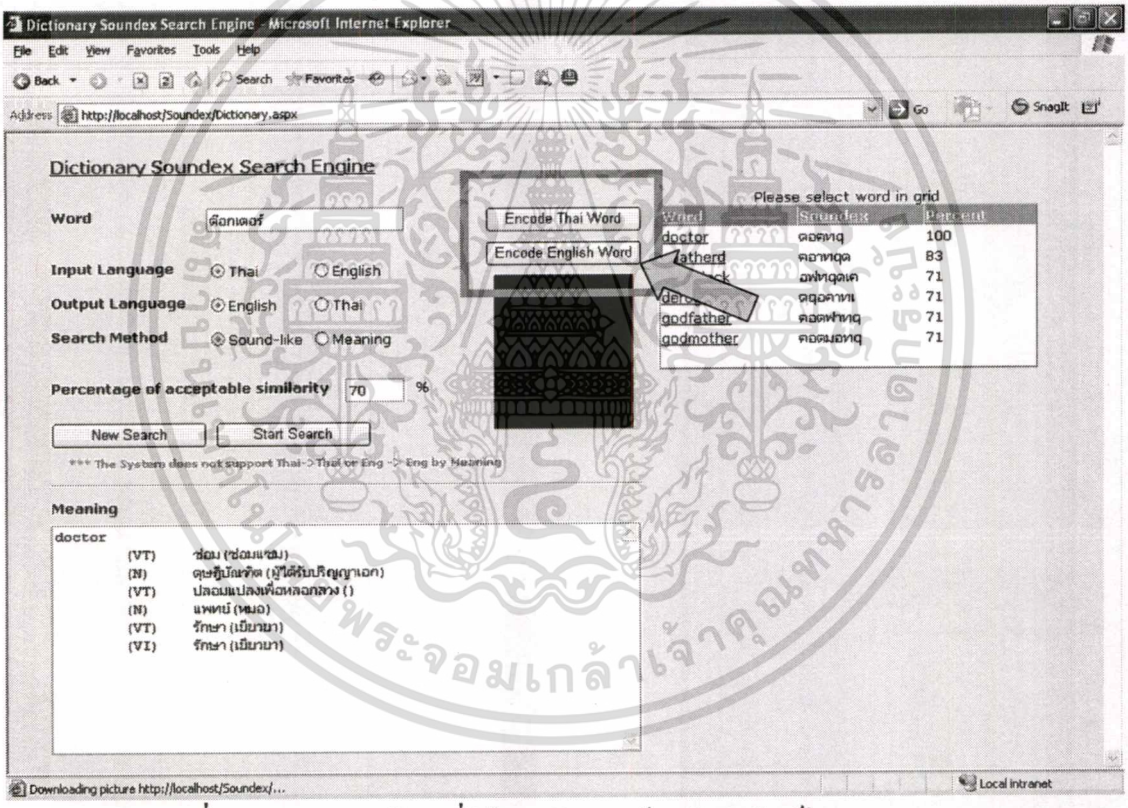
รูปที่ 5.1 หน้าจอการทำงานหลักของระบบ

โดยการทำงานของหน้าจอหลักแบ่งออกเป็น 2 ส่วน ได้แก่

1. การเข้ารหัสคำศัพท์ในพจนานุกรม
2. การค้นคืนคำศัพท์และแปลความหมายจากพจนานุกรม

5.3.1 การเข้ารหัสคำศัพท์ในพจนานุกรม

หน้าจอในส่วนนี้หน้าที่ในการนำคำจากฐานข้อมูลพจนานุกรมมาทำการเข้ารหัสตามอัลกอริทึมที่กำหนดไว้ เพื่อเตรียมพร้อมสำหรับการใช้งานระบบการค้นคืนคำศัพท์และแปลความหมายในขั้นตอนต่อไป หน้าจอเข้ารหัสคำศัพท์ในพจนานุกรมได้แสดงไว้ดังรูปที่ 5.2



รูปที่ 5.2 ส่วนของหน้าจอที่ใช้สำหรับการเข้ารหัสคำศัพท์ในพจนานุกรม

5.3.1.1 หน้าที่ของส่วนประกอบในหน้าจอการเข้ารหัสคำในพจนานุกรม

คำอธิบายหน้าที่ของแต่ละส่วนประกอบในหน้าจอการทำงานหลัก ในส่วนของการเข้ารหัสคำในฐานข้อมูลพจนานุกรม แสดงไว้ในตารางที่ 5.1

ตารางที่ 5.1 หน้าที่การทำงานในแต่ละส่วนประกอบของหน้าจอสำหรับการเข้ารหัสคำในพจนานุกรม

ลำดับที่	ส่วนประกอบของหน้าจอ	หน้าที่
1	ปุ่ม Encode Thai Word	เข้ารหัสคำศัพท์ภาษาไทย
2	ปุ่ม Encode English Word	เข้ารหัสคำศัพท์ภาษาอังกฤษ

5.3.1.2 ขั้นตอนการทำงานของการเข้ารหัสคำในพจนานุกรม

กรณีที่ผู้ใช้ต้องการเข้ารหัสคำศัพท์ภาษาไทยในฐานข้อมูลพจนานุกรม

1. ผู้ใช้กดปุ่ม “Encode Thai Word”
2. ระบบทำการดึงข้อมูลคำศัพท์ภาษาไทยจากฐานข้อมูลพจนานุกรม
3. ทำการเข้ารหัสคำศัพท์ภาษาไทย
4. บันทึกคำศัพท์ที่เข้ารหัสแล้วกลับไปยังฐานข้อมูล

กรณีที่ผู้ใช้ต้องการเข้ารหัสคำศัพท์ภาษาอังกฤษในฐานข้อมูลพจนานุกรม

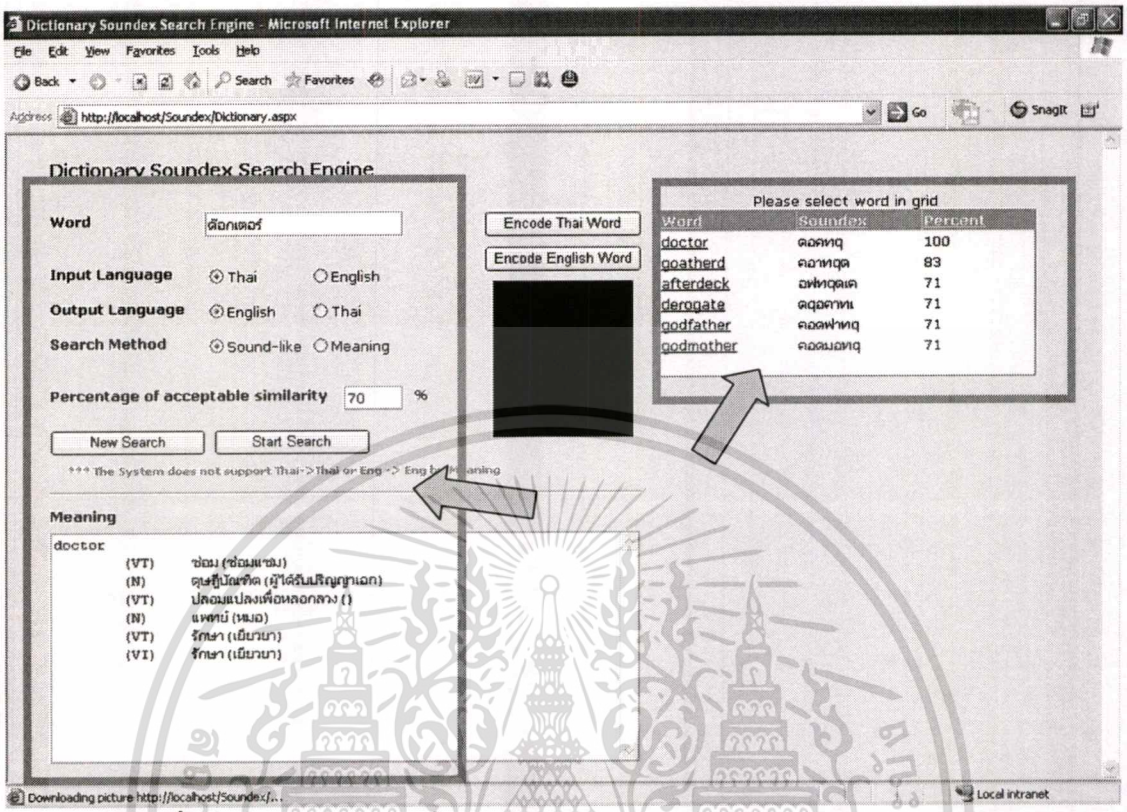
1. ผู้ใช้กดปุ่ม “Encode English Word”
2. ระบบทำการดึงข้อมูลคำศัพท์ภาษาอังกฤษจากฐานข้อมูลพจนานุกรม
3. ทำการเข้ารหัสคำศัพท์ภาษาอังกฤษ
4. บันทึกคำศัพท์ที่เข้ารหัสแล้วกลับไปยังฐานข้อมูล

5.3.2 การค้นคืนคำศัพท์และแปลความหมายจากพจนานุกรม

หน้าจอนี้มีหน้าที่ในการค้นหาคำศัพท์ในฐานข้อมูลพจนานุกรมที่มีคำอ่านคล้ายกับคำข้อคำถาม โดยจะแสดงรายการคำศัพท์ที่ผ่านเกณฑ์ออกมา เพื่อให้ผู้ใช้เลือกคำศัพท์ที่ต้องการทราบความหมาย ระบบจะแปลความหมายของคำศัพท์กลับมาแสดงผลทางหน้าจอ หน้าจอของระบบการค้นคืนคำศัพท์และแปลความหมาย แสดงไว้ดังรูปที่ 5.3

5.3.2.1 หน้าที่ของส่วนประกอบในหน้าจอการค้นคืนคำศัพท์และแปลความหมาย

คำอธิบายหน้าที่ของแต่ละส่วนประกอบในหน้าจอการทำงานหลัก ในส่วนของการค้นคืนคำศัพท์และแปลความหมาย แสดงไว้ในตารางที่ 5.2



รูปที่ 5.3 ส่วนของหน้าจอที่ใช้สำหรับการค้นคืนคำศัพท์และแปลความหมาย

ตารางที่ 5.2 หน้าที่การทำงานในแต่ละส่วนประกอบของหน้าจอสำหรับการค้นคืนคำศัพท์และแปลความหมาย

ลำดับที่	ส่วนประกอบของหน้าจอ	หน้าที่
1	Word	กรอกคำอ่านของคำที่ต้องการค้นหา (คำอ่านคำทับศัพท์ หรือคำพ้องเสียง)
2	Input Language	เลือกภาษาของคำที่เป็นอินพุต
3	Output Language	เลือกภาษาของคำที่เป็นเอาท์พุต
4	Search Method • Sound-like • Meaning	เลือกวิธีในการค้นหาคำศัพท์จากพจนานุกรม • ให้ค้นหาจากคำที่อ่านออกเสียงเหมือนกัน • ให้ค้นหาจากคำที่มีความหมายตรงกัน
5	Percentage of similarity acceptable	กำหนดค่าที่จะใช้เป็นเกณฑ์ตัดสินว่าคำที่ถูกนำมาเปรียบเทียบกันเป็นคำเดียวกันหรือไม่ โดยค่าความเหมือนคิดเป็นเปอร์เซ็นต์ ตั้งแต่

		0 ถึง 100 เปอร์เซ็นต์ (100 = เหมือนกันทุกประการ)
6	New Search	ใช้เพื่อล้างค่าตัวแปรทั้งหมด และเริ่มต้นขั้นตอนที่ 1 ใหม่อีกครั้ง
7	Start Search	เริ่มกระบวนการเปรียบเทียบและค้นหาคำจากพจนานุกรมที่ตรงกับเงื่อนไขที่ได้กำหนดไว้แล้วนำผลลัพธ์ที่ได้เก็บลงในกริด
8	Please select word in Grid	คลิกเลือกคำศัพท์จากกริด เพื่ออธิบายความหมายและรายละเอียดเพิ่มเติมของคำศัพท์คำนั้นในช่อง “Meaning” ทางด้านล่างของหน้าจอ
9	Meaning	แสดงความหมายของคำศัพท์ที่ถูกเลือกจากกริด

5.3.2.2 ขั้นตอนการทำงานของหน้าจอสำหรับการค้นคืนคำศัพท์และแปลความหมาย

1. ผู้ใช้กรอกคำอ่านของคำที่ต้องการค้นหา (คำอ่าน คำทับศัพท์ หรือคำพ้องเสียง) ในช่อง “Word”
2. ผู้ใช้เลือกภาษาของคำที่เป็นอินพุต เช่น ถ้าคำที่ผู้ใช้กรอกในขั้นตอนแรกเป็นภาษาไทย ให้เลือก “Thai”
3. ผู้ใช้เลือกภาษาของคำที่เป็นเอาท์พุต เช่น ถ้าต้องการให้ได้ผลลัพธ์เป็นคำศัพท์ภาษาอังกฤษ เลือก “English”
4. ผู้ใช้เลือกวิธีในการค้นหาคำศัพท์ในพจนานุกรม เช่น ถ้าต้องการค้นหาจากคำที่อ่านออกเสียงเหมือนกัน ให้เลือก “Sound-like” หรือ “Meaning” กรณีที่ต้องการค้นหาจากความหมาย
5. ผู้ใช้กำหนดค่าที่จะใช้เป็นเกณฑ์ตัดสิน โดยใส่ค่าความเหมือนคิดเป็นเปอร์เซ็นต์ มีค่าตั้งแต่ 0 ถึง 100 เปอร์เซ็นต์ (ค่าตั้งต้นมีค่าเท่ากับ 70 เปอร์เซ็นต์)
6. ผู้ใช้กดปุ่ม “Start Search” เพื่อเริ่มกระบวนการเปรียบเทียบและค้นหาคำจากพจนานุกรมที่ตรงกับเงื่อนไขที่ได้กำหนดไว้
7. คำศัพท์ที่ผ่านเกณฑ์จะแสดงไว้ในตารางกริด ผู้ใช้สามารถกดคลิกเลือกคำศัพท์จากกริดเพื่ออธิบายความหมายและรายละเอียดเพิ่มเติมของคำศัพท์คำนั้นในช่อง “Meaning” ด้านล่างของหน้าจอ
8. หากต้องการค้นหาคำศัพท์คำอื่น ให้กดปุ่ม “New Search” เพื่อกลับไปยังขั้นตอนแรก

บทที่ 6
บทสรุป

6.1 สรุปโครงการ

โครงการนี้ได้ศึกษาความเป็นไปได้ในการพัฒนาระบบพจนานุกรมให้สามารถใช้การค้าหาด้วยคำพ้องเสียง ทำให้เพิ่มความสามารถในการค้นคืนข้อมูลคำศัพท์ข้ามภาษาไทย-อังกฤษ โดยพัฒนาระบบในรูปแบบของเว็บแอพพลิเคชั่น ทำให้สามารถให้บริการผ่านเครื่อข่ายอินเทอร์เน็ตได้จากที่ต่าง ๆ ทั่วโลก แล้วได้นำเทคโนโลยีที่ทันสมัยเช่น การนำ Microsoft Visual Studio .NET 2005 และ Microsoft SQL Server 2005 มาใช้พัฒนาระบบ ทำให้สามารถทำการพัฒนาระบบได้อย่างรวดเร็วและมีประสิทธิภาพ

ระบบที่พัฒนาสามารถรองรับการใช้งานได้หลายรูปแบบเช่น การค้นหาคำศัพท์ข้ามภาษาไทย-อังกฤษ อังกฤษ-ไทย ไทย-ไทย และ อังกฤษ-อังกฤษ หรือการค้นหาโดยใช้ความหมายในการค้นหาข้อมูลคำศัพท์ได้เหมือนพจนานุกรมทั่วไป

6.2 ผลลัพธ์จากการพัฒนาระบบ

หลังจากที่ได้ทดสอบการทำงานของระบบพบว่า ระบบสามารถค้นคืนข้อมูลคำศัพท์ข้ามภาษาไทย-อังกฤษได้อย่างถูกต้องประมาณ 60-70% เนื่องจากการเข้ารหัสคำยังเน้นที่รูปมากกว่าการเน้นที่เสียงของคำ นอกจากนั้น ยังพบคำบางคำที่รูปและเสียงอ่านต่างกัน เช่น คำว่า “เรสเตอรอง” หรือ “restaurant” ในภาษาอังกฤษ เป็นต้น นอกจากนั้น การเข้ารหัสคำศัพท์โดยใช้ทฤษฎีต่าง ๆ ที่ผู้พัฒนาได้ศึกษามา พบว่ายังไม่สามารถทำงานได้ดีเท่าที่ควร ดังนั้น จึงมีหลายคำที่ผู้พัฒนาจำเป็นต้องบอกคำอ่านล่วงหน้ากับระบบเพื่อลดความผิดพลาดให้น้อยที่สุด

6.3 ข้อเสนอแนะและแนวทางการพัฒนาระบบ

1. ระบบการเข้ารหัสคำนี้ยังสามารถพัฒนาอัลกอริทึมต่อไปได้ เชื่อมั่นว่าระบบจะสามารถค้นคืนข้ามภาษาได้ถูกต้องแม่นยำมากขึ้น
2. เราสามารถนำระบบการค้นคืนข้ามภาษาไปประยุกต์ใช้งานกับแอพพลิเคชั่นในหลายรูปแบบ เช่น ระบบการค้นคืนข้อมูลจากอินเทอร์เน็ต ระบบค้นหาข้อมูลหนังสือตามห้องสมุด หรือร้านหนังสือ ซึ่งล้วนแล้วแต่มีคำทับศัพท์ จำนวนมากที่ระบบในปัจจุบันไม่สามารถค้นคืนสิ่งเหล่านี้ได้

บรรณานุกรม

จันทร์เพ็ญ โวหารสุนทร. 2530. **"การศึกษาการใช้อักษรโรมันแทนอักษรไทย."** ปริญญานิพนธ์, คณะอักษรศาสตร์ มหาวิทยาลัยศรีนครินทร์วิโรฒ ประสานมิตร.

ฉัตรชัย สุขสอาด. 2545. **Web Services ABC.** [Online]. เข้าถึงได้จาก: http://www.wsiam.com/document/abcwebservices/webservicesabc.jsp.

ทัศนวรรณ ศูนย์กลาง และคณะ. 2543. "การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษเพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก." หน้า 158-165. ใน **การประชุมวิชาการทางด้านวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติครั้งที่ 4.** กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล. 2542. **ขั้นตอนวิธีการเข้ารหัสและการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษ.** [Online]. เข้าถึงได้จาก : http://www.cp.eng.chula.ac.th/~somchai/spj/papers /ThaiText/NCSEC99_2C.pdf.

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 2537. **โปรแกรมสืบค้นข้อมูล สรรสาร.** [Online]. เข้าถึงได้จาก: http://www.sansarn.com.

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 2546. **เล็กซิตรอน ดิกชั่นนารี.** [Online]. เข้าถึงได้จาก: http://lexitron.nectec.or.th.

Holmes, D. and McCabe, C. 2002. **Improving Precision and Recall for Soundex Retrieval.** [Online]. Available: http://ir.iit.edu/publications/downloads/IEEESoundexV5.pdf.

Suwanvisat, Prayut and Prasitjutrakul, Somchai. 1998. **Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique.** [Online]. Available: http://www.cp.eng.chula.ac.th/~somchai/spj/papers/ThaiText/ncsec98-clir.pdf.

MSDN. 2005. **Microsoft Developer Tools Roadmap 2004-2005**. [Online]. Available: http://msdn.microsoft.com/vstudio/productinfo/roadmap.aspx.

ประวัติผู้เขียน

ชื่อผู้เขียน นายธนศักดิ์ ไชยะกุล

สถานที่เกิด จังหวัดกรุงเทพมหานคร

ประวัติการศึกษา

ระดับประถมศึกษา	โรงเรียนอัสสัมชัญ สำโรง จังหวัดสมุทรปราการ
ระดับมัธยมศึกษาตอนต้น	โรงเรียนอัสสัมชัญ สำโรง จังหวัดสมุทรปราการ
ระดับมัธยมศึกษาตอนปลาย	โรงเรียนวัดสุทธิวราราม ยานนาวา จังหวัดกรุงเทพมหานคร
ระดับอุดมศึกษา	วิทยาศาสตร์บัณฑิต (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยกรุงเทพ

ประวัติการทำงาน

เมษายน 2541 – พฤษภาคม 2547	ศูนย์คอมพิวเตอร์ มหาวิทยาลัยกรุงเทพ ตำแหน่งอาจารย์ประจำศูนย์คอมพิวเตอร์
พฤษภาคม 2547 - ปัจจุบัน	บริษัท ดีเอสที อินเทอร์เนชั่นแนล (ประเทศไทย) จำกัด ตำแหน่งที่ปรึกษาทางเทคนิค