บทที่ 8

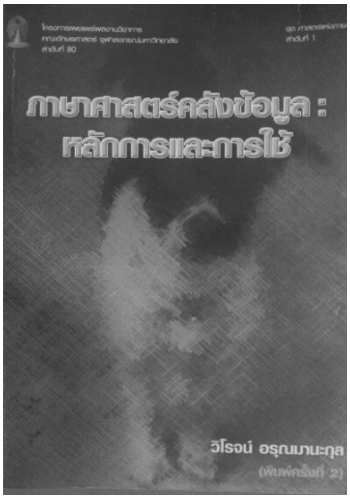
ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้

พิสุทธิ์พงศ์ เอ็นดู
Pisutpong Endoo

บทที่ 8

ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้[1]

พิสุทธิ์พงศ์ เอ็นดู[2]
Pisutpong Endoo



หากผู้อ่านสนใจที่จะศึกษาค้นคว้าหาความรู้หรือทำงานเกี่ยวกับภาษาศาสตร์ วิชาที่ว่าด้วยการศึกษาเกี่ยวกับภาษาทั่วๆไปของมนุษย์ตามแนววิทยาศาสตร์ โดยมีการนำเทคโนโลยีสมัยใหม่ทางคอมพิวเตอร์มาประยุกต์ใช้กับการศึกษาภาษาศาสตร์ จนแตกเป็นแขนงขึ้นมาในสาขาวิชาภาษาศาสตร์ที่เรียกขานกันคือ ภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) และในสาขาวิชาวิทยาการคอมพิวเตอร์ที่นิยมเรียกกันคือ การประมวลภาษาธรรมชาติ (Natural language processing) เป็นการผนวกกันของทั้งสองสาขาวิชานี้เข้าด้วยกัน จนเป็นที่มาของการนำเสนอเนื้อหาสาระและประเด็นที่สำคัญของหนังสือเล่มนี้ “ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้” เขียนโดยอาจารย์วิโรจน์ อรุณมานกุล ดังนั้นเพื่อๆให้อ่านและผู้สนใจทั่วไปทราบถึงรายละเอียดเกี่ยวกับหนังสือเล่มนี้ ในการนำเสนอเนื้อหานี้จะแบ่งออกเป็นหัวข้อหลักๆประกอบไปด้วย ความรู้เบื้องต้นเกี่ยวกับหนังสือ เนื้อหาในหนังสือโดยสรุป จุดเด่นของหนังสือ จุดที่ควรเพิ่มเติมในหนังสือและข้อเสนอแนะสำหรับผู้อ่าน

[1] วิโรจน์ อรุณมานะกุล. (2553). ภาษาศาสตร์คลังข้อมูล: หลักการและการใช้. (พิมพ์ครั้งที่ 2). กรุงเทพฯ : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 387หน้า.

[2] อาจารย์ประจำสาขาวิชาภาษาศาสตร์ คณะเทคโนโลยีการจัดการ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ [pisutpong123@hotmail.com]

ความรู้เบื้องต้นเกี่ยวกับหนังสือ

"ภาษาศาสตร์คลังข้อมูล: หลักการและการใช้"เขียนโดยอาจารย์วิโรจน์ อรุณมานะกุล สำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิศวคอมพิวเตอร์ จากจุฬาลงกรณมหาวิทยาลัยในปี พ.ศ.2533 ปริญญาโท อักษรศาสตรมหาบัณฑิต สาขาภาษาศาสตร์ จากจุฬาลงกรณ์ราชวิทยาลัยในปี พ.ศ.2539 และปริญญาเอก สาขาภาษาศาสตร์คอมพิวเตอร์ จากมหาวิทยาลัยจอร์จทาวน์ ประเทศสหรัฐอเมริกา ในปี พ.ศ.2542 ประสบการณ์เคยทำงานในตำแหน่งวิศวกรควบคุมระบบ (System Engineer) ดูแลรับผิดชอบระบบบริการสอบถามเลขหมาย 13 ขององค์การโทรศัพท์แห่งประเทศไทยระหว่างปี พ.ศ.2529-2530 และทำงานเป็นเจ้าหน้าที่ระบบงานคอมพิวเตอร์ สถาบันบริการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยระหว่างปี พ.ศ. 2531-2533 ในปัจจุบันเป็นอาจารย์ประจำภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ชีวิตครอบครัวสมรสกับอาจารย์วิไลวรรณ อรุณมานะกุล อาจารย์ประจำสถาบันภาษา มหาวิทยาลัยธรรมศาสตร์ มีผลงานด้านวิชาการนั้นมีผลงานการเขียนตำราวิชาการ การวิจัย การนำเสนอผลงานวิชาการทั้งในประเทศและต่างประเทศ รวมกันไม่น้อยกว่า 50 เรื่อง

หนังสือ "ภาษาศาสตร์คลังข้อมูล: หลักการและการใช้" เป็นหนังสือที่ให้ความรู้พื้นฐานต่างๆเกี่ยวข้องกับคลังข้อมูลภาษาเพื่อให้ผู้อ่านได้เข้าใจและเห็นความสำคัญของการนำคลังข้อมูลภาษาไปใช้ประโยชน์ในการศึกษาด้านต่างๆ โดยมีข้อมูลพื้นฐานที่พัฒนาจากรายงานการวิจัยเรื่อง "แนวทางการพัฒนาคลังข้อมูลภาษา:กรณีศึกษาจากการสร้างคลังข้อมูลภาษาอังกฤษ" ซึ่งเป็นโครงการวิจัยของผู้เขียนเอง โดยมีวัตถุประสงค์เพื่อศึกษากลวิธีที่ใช้ในการสร้างคลังข้อมูลทางภาษา ผู้เขียนได้ทำการสำรวจคลังข้อมูลภาษาจากแหล่งต่างๆ จากนั้นได้นำมาวิเคราะห์ถึงลักษณะความแตกต่างรวมทั้งปัจจัยที่กำหนดลักษณะของการจัดวางโครงสร้างข้อมูลในแต่ละแบบ นอกเหนือจากนี้แล้วผู้เขียนได้ดำเนินการทดลองสร้างคลังข้อมูลภาษาขึ้น โดยคัดเลือกเอกสารอิเล็กทรอนิกส์ภาษาอังกฤษที่สามารถสืบค้นได้ทางอินเตอร์เน็ตซึ่งไม่มีข้อจำกัดเรื่องลิขสิทธิ์ โดยข้อมูลที่รวบรวมมาได้จัดทำเป็นคลังข้อมูลภาษาอังกฤษเพื่อใช้ประโยชน์สำหรับการเรียนการสอนและงานวิจัยในมหาวิทยาลัย ผลของการศึกษาของผู้เขียนทั้งหมดนี้เองได้ถูกนำมาปรับปรุงเป็นข้อมูลงานเขียนในหนังสือเล่มนี้ โดย

ผู้เขียนได้เน้นถึงการศึกษาภาษาในแง่มุมต่างๆซึ่งมีข้อมูลการศึกษาภาษาเป็นจำนวนมากจากแหล่งต่างๆ โดยข้อมูลที่รวบรวมเป็นจำนวนมากนี้ถือเป็นคลังข้อมูลทางภาษาที่สามารถนำไปใช้ประโยชน์ได้ทั้งในแง่ของการเรียนการสอนภาษา การวิจัยทางภาษา การแปล การจัดทำพจนานุกรม และการประมวลผลภาษาธรรมชาติ

เนื้อหาในหนังสือโดยสรุป

เนื้อหาในหนังสือเล่มนี้ผู้เขียนได้แบ่งเนื้อหาสำหรับการนำเสนอออกเป็น 13 บท โดยในแต่ละบทนั้นก็มีประเด็นย่อยมีความสัมพันธ์และสอดคล้องกับเนื้อหาในแต่ละบทผู้เขียนได้หยิบยกประเด็นสำคัญๆมานำเสนอพร้อมยกตัวอย่างเพื่อให้ผู้อ่านมองเห็นภาพและสามารถจินตนาการติดตามเนื้อหาร่วมกับผู้เขียนได้ สำหรับเนื้อหาทั้งหมดนั้นสามารถสรุปได้ออกเป็น 3 ตอนคือ ตอนที่ 1 การนำคลังข้อมูลกับการนำไปใช้ ตอนที่ 2 การออกแบบและการสร้างคลังข้อมูลและตอนที่ 3 โปรแกรมและการใช้เว็บเพื่อการศึกษาคลังข้อมูล

ตอนที่ 1 การนำคลังข้อมูลกับการนำไปใช้

ในหนังสือเล่มนี้จากการประมวลเนื้อหาพบว่าบทที่ 1 ถึงบทที่ 7 มีเนื้อหาเกี่ยวข้องกับการนำคลังข้อมูลกับการนำไปใช้ โดยในรายละเอียดในแต่ละบทประกอบไปด้วยเนื้อหาโดยสรุปดังนี้

บทที่ 1 คลังข้อมูลทางภาษา ในบทนี้ผู้เขียนนำเสนอความรู้เบื้องต้นเกี่ยวกับคลังข้อมูลคืออะไร เพราะเหตุใดจึงได้รับความสนใจจากนักภาษาศาสตร์และนักคอมพิวเตอร์เป็นจำนวนมาก โดยผู้เขียนได้แบ่งเนื้อหาของการนำเสนอในบทนี้ออกเป็น 8 หัวข้อได้แก่ 1.1) คลังข้อมูลภาษาคืออะไร 1.2) ความเป็นมาของภาษาศาสตร์คลังข้อมูล 1.3) ขอบเขตของภาษาศาสตร์คลังข้อมูล 1.4) พัฒนาการของคลังข้อมูลภาษา 1.5) ประเภทของคลังข้อมูลภาษา 1.6) ความหมายของ CORPUS และ ARCHIVE 1.7) คำนวณภาพขนาดของคลังข้อมูล และ 1.8) ปัจจุบันและอนาคตของภาษาศาสตร์คลังข้อมูล สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 1 นั้นคือผู้เขียนได้ยกคำนิยามรวมทั้งอธิบายถึงลักษณะของคลังข้อมูลภาษาที่เป็นภาษาอังกฤษโดยมีการอ้างถึงผู้เขียนและแหล่งที่มาประกอบไว้เพื่อยืนยันถึงข้อมูลที่ผู้เขียนได้นำมาอ้างอิงและกล่าวถึงควบคู่กับภาษาไทยที่ผู้เขียนได้แปลและอ้างอิงไว้

บทที่ 2 คลังข้อมูลกับการเรียนการสอนภาษา ในบทนี้ผู้เขียนกล่าวถึงประโยชน์ของคลังข้อมูลที่มีต่องานด้าน ได้แก่ ด้านการเรียนการสอน งานด้านการศึกษาวิจัย

ทางภาษาศาสตร์ งานด้านการแปล และงานด้านทางด้านการประมวลผลภาษาธรรมชาติ เพื่อให้ผู้อ่านได้มีความรู้ความเข้าใจว่าคลังข้อมูลภาษาคืออะไรและมีประโยชน์อย่างไรบ้าง ในเนื้อหาของบทนี้ผู้เขียนได้จำแนกการนำเสนอเนื้อหาออกเป็นหัวข้อย่อยๆประกอบไปด้วย 2.1) บทนำ 2.2) การนำคลังข้อมูลภาษามาใช้เป็นหลักในการสอนภาษา และ 2.3) การประยุกต์ใช้คลังข้อมูลภาษาในการสอนภาษา สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 2 นั้นได้แก่ การนำเสนอตัวอย่างของแบบฝึกหัดของการนำคลังข้อมูลมาเพื่อใช้กับการเรียนการสอน โดยตัวอย่างแบบฝึกหัดที่นำมาเสนอนั้นมีทั้งที่เป็นการใช้คลังข้อมูลเพื่อการศึกษาส่วนของคำพูด (parts of speech) และหลักไวยากรณ์ภาษาอังกฤษ เป็นต้น

บทที่ 3 คลังข้อมูลกับการวิจัยทางภาษา ในบทนี้ผู้เขียนได้อธิบายถึงการทำวิจัยทางภาษาโดยใช้ข้อมูลที่มีอยู่ในคลังข้อมูลที่มีอยู่เป็นจำนวนมาก ในบทนี้ผู้เขียนได้เขียนชี้ประเด็นให้ผู้อ่านเห็นถึงแนวทางการทำวิจัยทางภาษาโดยใช้ข้อมูลที่มีอยู่ในคลังข้อมูลว่าสามารถนำมาเรียนรู้และทำงานวิจัยที่เกี่ยวข้องการศึกษาทางภาษาได้ โดยผู้เขียนได้แบ่งหัวข้อสำหรับการอธิบายดังนี้ 3.1) บทนำ 3.2) การศึกษาระดับคำ 3.3) การศึกษาระดับโครงสร้าง 3.4) การศึกษาด้าน ESP (English for Specific Purpose) และการแปรตามทำเนียบภาษาและ 3.5) หัวข้อในการวิจัยทางภาษาโดยใช้คลังข้อมูล สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 3 นั้นได้แก่ ตัวอย่างหัวข้อการวิจัย รายชื่อหรือหน่วยงานผู้วิจัย พร้อมทั้งผลของการศึกษาวิจัยซึ่งเกี่ยวข้องกับคลังข้อมูล ผู้วิจัยได้นำเสนอตัวอย่างเป็นจำนวนมากทำให้ผู้อ่านหรือผู้ศึกษาสามารถนำมาใช้เป็นแนวทางการศึกษาและทำวิจัยในอนาคตได้

บทที่ 4 คลังข้อมูลกับการแปล ในบทนี้ผู้เขียนเริ่มต้นบรรยายถึงความสำคัญและประโยชน์ของคลังข้อมูลที่จะถูกนำมาเป็นทรัพยากรในการแปลจากภาษาหนึ่งสู่อีกภาษาหนึ่ง โดยเฉพาะทรัพยากรที่เป็นคลังข้อมูลที่สำคัญในการนำไปใช้เป็นเครื่องมือในการแปลภาษา ยกตัวอย่างเช่น พจนานุกรมเฉพาะทาง พจนานุกรมสองภาษาและสารานุกรมอื่นๆ นอกเหนือจากนี้แล้วในบทนี้ผู้เขียนได้นำเสนอแนวคิดของ Baker ในปี ค.ศ.1993 ซึ่งมีความเชื่อว่าแนวโน้มของศาสตร์การแปลกำลังมีการพัฒนาการและเปลี่ยนแปลงไปในทางที่ดีขึ้นกว่าเดิม สืบเนื่องมาจากการมีคลังข้อมูลของภาษาต้นฉบับและภาษาแปลมากขึ้น ประเด็นนี้จะส่งผลทำให้มีผู้สนใจในการแปลและมีความศึกษาบทแปลได้ชัดเจนมากขึ้นกว่าเดิม นอกเหนือจากนั้นผู้เขียนชี้ให้เห็นถึงประโยชน์ของคลังข้อมูลที่มีต่องานแปลซึ่งมีความสำคัญอย่างมากในฐานะเป็นเครื่อง

มือในการช่วยให้ผู้แปลให้เกิดความง่ายและสะดวกสบายมากยิ่งขึ้น ในบทนี้ผู้เขียนได้แบ่งหัวข้อเพื่อให้ง่ายต่อการอธิบายและยกตัวอย่างประกอบในแต่ละข้อดังนี้ 4.1) ความเป็นมาของศาสตร์ 4.2) ทฤษฎีระบบหลายขั้ว 4.4) แนวคิดเรื่องรูปแบบ 4.5) คลังข้อมูลภาษากับศาสตร์การแปลแบบพรรณนา 4.6) ศาสตร์การแปลและการวิเคราะห์เปรียบต่าง 4.7) การศึกษาลักษณะเฉพาะของการและ 4.8) การใช้คลังข้อมูลภาษาเพื่อช่วยในการแปล สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 4 นั้นได้แก่ การนำเสนอตัวอย่างของการศึกษาและการทดลองการนำคลังข้อมูลมาใช้กับงานแปล ผู้เขียนได้ยกตัวอย่างงานศึกษาและงานวิจัยของนักวิชาการในต่างประเทศเป็นจำนวนมากทั้งที่เป็นชาวเอเชียและชาวยุโรปซึ่งมีผลงานการทำวิจัยเกี่ยวกับการนำคลังข้อมูลมาใช้สำหรับงานแปล การนำเสนอผลงานการวิจัยเหล่านี้ผู้เขียนยังได้อธิบายถึงวิธีการศึกษาและเครื่องมือเพื่อการศึกษาและการทำวิจัยของนักวิชาเหล่านั้นมาประกอบด้วย การอธิบายในลักษณะนี้ช่วยทำให้ผู้อ่านเกิดความเข้าใจและสามารถเรียนรู้กระบวนการทำวิจัยในงานคลังข้อมูลกับงานแปลได้ง่ายมากยิ่งขึ้น

บทที่ 5 คลังข้อมูลกับการทำพจนานุกรม ในบทนี้ผู้เขียนได้นำเสนอเกี่ยวกับข้อมูลซึ่งถูกรวบรวมและจัดเป็นหมวดหมู่อย่างเป็นระบบและแบบแผนในคลังข้อมูลว่าสามารถนำมาใช้ประกอบการจัดทำพจนานุกรมได้อย่างไร ทั้งนี้ผู้เขียนได้ชี้ประเด็นให้ผู้อ่านเห็นว่าคำศัพท์ในคลังข้อมูลนั้นมีความสำคัญและมีความเกี่ยวเนื่องกับการดำเนินการจัดทำพจนานุกรมอย่างไรบ้างและโดยวิธีการใด เพื่อให้เนื้อหามีความกระชับและชัดเจน ผู้เขียนได้แบ่งหัวสำหรับการอธิบายออกเป็น 5.1) บทนำ 5.2) พจนวิทยาและการทำพจนานุกรม 5.3) การทำพจนานุกรมแบบใช้คลังข้อมูลภาษา 5.4) อรรถศาสตร์คำศัพท์ 5.5) เครือข่ายคำ 5.6) ศัพท์วิทยาและการทำประมวลศัพท์ และ 5.7) การสรุปในภาพของการการจัดทำคลังข้อมูลกับการทำพจนานุกรมดังกล่าว สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 5 นั้นได้แก่ การนำเสนอภาพตัวอย่างการปรากฏร่วมของคำต่างๆ ซึ่งมีการแสดงให้เห็นถึงการแยกตามประเภทของคำเช่น คำนาม คำคุณศัพท์ เป็นต้น นอกเหนือจากนี้แล้วผู้เขียนได้ยังนำเสนอภาพตารางการแจกแจงข้อมูลของคำภาษาอังกฤษในชนิดต่างๆมาประกอบการอธิบายในหัวข้อแต่ละหัวข้อไว้ด้วยทำให้ผู้อ่านนั้นสามารถมองเห็นภาพและเข้าใจเนื้อหาขั้นตอนการจัดทำพจนานุกรมได้ง่ายมากยิ่งขึ้น

บทที่ 6 คลังข้อมูลกับการประมวลผลทางภาษา เนื้อหาในบทนี้ผู้เขียนนำเสนอเกี่ยวกับวิธีการทางสถิติ ข้อมูลทางสถิติที่จำเป็นต่อการประมวลผลทางภาษา และวิธี

การใช้สถิติเพื่อช่วยหาคำตอบที่มักปรากฏใช้ร่วมกันซึ่งเป็นการใช้สถิติร่วมกันกับคลังข้อมูลเพื่อการค้นหาคำตอบทางภาษาในอีกลักษณะหนึ่ง ในเนื้อหาของบทนี้ผู้เขียนได้แบ่งหัวข้อประกอบไปด้วย 6.1) บทนำ 6.2) ความสัมพันธ์ระหว่างภาษาศาสตร์และสถิติและ 6.3) การใช้สถิติในการแก้ปัญหาความกำกวม และ 6.4) สรุป สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 6 นั้นได้แก่ การนำเสนอตารางการประมวลผลทางภาษาต่างๆ ซึ่งมีจำนวน 4 ตารางและรูปภาพของผัง Markov และรูปแสดงความเป็นไปได้ของลำดับหมวดคำ ตารางและรูปภาพที่ผู้เขียนนำเสนอในบทนี้นั้น ผู้เขียนได้อธิบายถึงการวิธีการศึกษาข้อมูลในตารางและรูปภาพอย่างละเอียด โดยตารางและรูปภาพที่นำมาเสนอนั้นก็ได้อ้างอิงถึงแหล่งที่มาประกอบไว้ด้วย

บทที่ 7 คลังข้อมูลภาษาผู้เรียน เนื้อหาในบทนี้ผู้เขียนได้กล่าวถึงคลังข้อมูลที่ผู้เรียนผู้สอนหรือผู้สนใจสามารถเลือกนำมาใช้เพื่อการเรียนหรือนำมาประยุกต์เพื่อการศึกษา ค้นคว้าและทำวิจัยเกี่ยวกับงานทางด้านภาษา เพื่อให้เนื้อหาในบทนี้มีความกระชับและง่ายต่อการทำความเข้าใจ ผู้เขียนได้แยกอธิบายออกเป็นหัวข้อย่อยประกอบไปด้วย 7.1) บทนำ 7.2) การใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียน 7.3) การพัฒนาคลังข้อมูลภาษาผู้เรียน 7.4) โครงการพัฒนาคลังข้อมูลภาษาผู้เรียนที่ต่างๆ 7.5) ตัวอย่างงานวิจัยที่ใช้คลังข้อมูลภาษาผู้เรียน 7.6) การกำกับข้อผิดและ 7.7) บทสรุป สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 7 นั้น ผู้เขียนได้นำเสนอภาคผนวก ICLE Learner Profile ในท้ายบทเพื่อใช้อ้างอิงการอธิบายเนื้อหาในบทนี้ประกอบไว้ด้วย การนำเสนอภาคผนวกกำกับไว้ท้ายบทดังกล่าวเป็นวิธีการที่ช่วยให้ผู้อ่านนั้นสามารถทำความเข้าใจไปพร้อมๆกับเนื้อหาที่ผู้เขียนได้อธิบายไว้ได้ง่ายมากยิ่งขึ้น

ตอนที่ 2 การออกแบบและการสร้างคลังข้อมูล มีเนื้อหาเกี่ยวข้องกับวิธีการออกแบบและการสร้างคลังข้อมูลอย่างละเอียด เนื้อหาในตอนที่ 2 ประกอบไปด้วยบทที่ 8 ถึงบทที่ 10 โดยเนื้อหาในแต่ละบทนั้นมีเนื้อหาสรุปดังนี้

บทที่ 8 การออกแบบและการสร้างคลังข้อมูลทางภาษา เนื้อหาในบทนี้ผู้เขียนกล่าวถึงวิธีการออกแบบและการสร้างคลังข้อมูลทางภาษา ผู้เขียนนำเข้าผู้อ่านเข้าสู่เนื้อหาโดยอธิบายถึงวิธีการตั้งแต่การเตรียมตัวด้วยวิธีการง่ายๆซึ่งผู้ออกแบบและสร้างคลังข้อมูลทางภาษาด้วยตัวเองนั้นจำเป็นต้องพิจารณาและทบทวนในประเด็นต่างๆซึ่งมีความสำคัญอย่างมากก่อนที่จะลงมือทำ โดยแบ่งหัวข้อสำหรับการอธิบายขั้นตอนการออกแบบและการสร้างคลังข้อมูลทางภาษาดังนี้ 8.1) ข้อควรพิจารณา 8.2) ขั้นตอนในการจัดสร้างคลังข้อมูลภาษาและ 8.3) ตัวอย่างการจัดสร้างคลังข้อมูลภาษา

บทที่ 9 การนำข้อมูลภาษาเข้าคอมพิวเตอร์ เนื้อหาในบทนี้ผู้เขียนกล่าวถึงวิธีการนำข้อมูลภาษาที่ได้รวบรวมบันทึกเอาไว้ในคอมพิวเตอร์ โดยมีการอธิบายลักษณะการนำข้อมูลภาษาบันทึกเอาไว้ในคอมพิวเตอร์เป็นข้อๆดังนี้ 9.1) การเก็บข้อมูลในคอมพิวเตอร์ 9.2) รหัสอักขระ รหัสอักขระภาษาต่างๆ ตัวอักษรภาษาไทยและรหัสยูนิโค้ด 9.3) ไฟล์ข้อมูลแบบต่างๆ 9.4) การแปลงไฟล์ข้อมูล 9.5) การใช้โปรแกรมรู้จำตัวอักษรและแสกนเนอร์ 9.6) การใช้โปรแกรม HTTrack ซึ่งมีตัวเลือกต่างๆในโปรแกรม รวมถึงข้อแนะนำในการใช้โปรแกรมดังกล่าวและ 9.7) พระราชบัญญัติลิขสิทธิ์ พ.ศ.2537

บทที่ 10 การกำกับข้อมูล เนื้อหาในบทนี้ผู้เขียนได้กล่าวถึงการกำกับข้อมูลคือการเพิ่มเติมข้อมูลอื่นๆนอกเหนือจากตัวข้อความเอกสารนั้นๆ ทั้งนี้อาจจะเป็นการบันทึกเครื่องหมายหรือสัญลักษณ์ต่างๆที่เขียนเพิ่มเติมในเอกสารเพื่อบอกว่าจะให้จัดพิมพ์เอกสารนั้นๆออกมาอย่างไรเมื่อมีการจัดส่งเรียงพิมพ์ โดยการกำกับข้อมูลนั้นถูกนำมาใช้ประโยชน์ในการวิเคราะห์ทางภาษาศาสตร์ เพราะคลังข้อมูลภาษาที่มีเพียงข้อความล้วนอย่างเดียวไม่สามารถมาใช้ประโยชน์ได้จำเป็นต้องมีการใส่ข้อมูลเพิ่มเติมเพื่อกำกับและควบคุมขอบเขตการศึกษา ในบทนี้ผู้เขียนได้แบ่งหัวข้ออธิบายประกอบไปด้วย 10.1) เหตุผลที่มีการกำกับข้อมูล 10.2) การกำกับข้อมูลคืออะไร 10.3) หน้าที่และบทบาทของ TEI 10.4) SGML และ XML ซึ่งประกอบไปด้วยองค์ประกอบของเอกสาร SGML/XML มี SGML Declaration, XML Declaration, Document instance, Document Type Definition (DTD) การกำหนดส่วนเฉพาะ (Marked Section) และตัวอย่างของแฟ้มข้อมูลที่กำกับด้วย SGMLและXML 10.5) แท็กพื้นฐานที่กำหนดโดย TEI มีโครงสร้างของเอกสารตามข้อกำหนด TEI การกำกับข้อมูลส่วนเนื้อความ การกำกับข้อมูลในคลังข้อมูลภาษา การกำกับคลังข้อมูลภาษาตามแนวทางของ TEI แยกออกเป็น การกำกับข้อมูลบริบทและการกำกับข้อมูลทางภาษาศาสตร์ และ 10.6) สรุป

สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 8-10 นั้นได้แก่รูปแบบการอธิบายขั้นตอนการออกแบบและการสร้างคลังข้อมูลแบบง่ายๆเป็นขั้นเป็นตอน การอธิบายลักษณะดังกล่าวนี้ทำให้ผู้อ่านสามารถทำเองได้แม้ว่าจะไม่มีพื้นฐานของการสร้างและออกแบบคลังข้อมูลทางภาษามาก่อน วิธีการอธิบายของผู้เขียนดังกล่าวนี้ถือได้ว่าเป็นเสน่ห์ที่น่าดึงดูดสำหรับผู้อ่านหนังสือวิชาการเล่มนี้อีกลักษณะหนึ่ง

ตอนที่ 3 โปรแกรมและการใช้เวปเพื่อการศึกษาคลังข้อมูล มีเนื้อหาเกี่ยวข้องกับการนำโปรแกรมและการนำเว็บต่างๆในอินเทอร์เน็ตมาประกอบการศึกษาคลังข้อมูล โดยเนื้อหาในตอนที่ 3 นั้นประกอบไปด้วยบทที่ 11-13 ดังนี้

บทที่ 11 โปรแกรมคอนคอร์แด๊นซ์ เนื้อหาในบทนี้ผู้เขียนได้กล่าวถึงข้อมูลเกี่ยวกับโปรแกรมคอนคอร์แด๊นซ์ซึ่งเป็นโปรแกรมที่ใช้สำหรับค้นหาคำที่ต้องการและจัดเรียงคำนั้นพร้อมบริบทที่ปรากฏโดยคำที่ต้องการนั้นจะถูกจัดเรียงไว้เป็นแนวตรงกันอยู่กลางหน้ากระดาษ เพื่อให้ผู้อ่านประวัติความเป็นมาและข้อมูลที่เกี่ยวข้องกับโปรแกรมคอนคอร์แด๊นซ์ให้ชัดเจนมากยิ่งขึ้น ในบทนี้ผู้เขียนได้แบ่งหัวข้อสำหรับการอธิบายประกอบไปด้วย 11.1) บทนำ 11.2) ประเภทของโปรแกรมคอนคอร์แด๊นซ์ 11.3) การค้นข้อมูล 11.4) การแสดงผล 11.5) ข้อจำกัดของโปรแกรมคอนคอร์ดแด๊นซ์ 11.6) ประโยชน์ของโปรแกรมคอร์นคอร์แด๊นซ์ ซึ่งประกอบไปด้วย ประโยชน์ด้านการวิจัยทางภาษาศาสตร์ ประโยชน์ด้านการสอนภาษาและประโยชน์ในงานประยุกต์อื่นๆ 11.7) ข้อควรระวัง 11.8) ตัวอย่างการใช้โปรแกรมคอนคอร์แด๊นซ์ ซึ่งประกอบไปด้วย การโหลดคลังข้อมูลเข้าโปรแกรม การหารายการความถี่ของคำ การค้นคำหลักพร้อมบริบท การหาคำปรากฏร่วม การค้นหากลุ่มคำ การค้นแบบซับซ้อน การแสดงตำแหน่งการปรากฏของคำ การหาคำสำคัญเมื่อเทียบกับคลังข้อมูลอ้างอิง การใช้ AntConce กับข้อมูล XML และการใช้ AntConc กับข้อมูลภาษาไทย 11.9) การใช้โปรแกรมคอนคอร์แด๊นซ์ภาษาไทยและ11.10) การใช้คลังข้อมูลภาษาไทยแห่งชาติ

บทที่ 12 คำปรากฏร่วมและหน่วยเสมือนวลี เนื้อหาในบทนี้ผู้เขียนได้กล่าวถึงคำศัพท์ที่ปรากฏหรือมักปรากฏใช้ร่วมกับคำหรือวลีอื่นจนถูกนำไปใช้อย่างแพร่หลายและเป็นเรื่องปรกติ อาทิเช่น look ปรากฏใช้ร่วมกับ after เมื่อรวมกันจะกลายเป็น look after แปลว่าดูแล การปรากฏร่วมของคำว่า look กับ after เมื่อถูกนำไปใช้จะได้ความหมายใหม่เป็นต้น ในบทนี้ผู้เขียนได้แบ่งหัวข้อสำหรับการอธิบายเนื้อหาประกอบไปด้วย 12.1) ความเป็นมา 12.2) ประเภทของคำปรากฏร่วม 12.3) เกณฑ์การพิจารณาคำปรากฏร่วม 12.4) การระบุหาคำปรากฏร่วม ซึ่งประกอบไปด้วย การใช้ความถี่ การใช้ค่าเฉลี่ยและความแปรปรวน การทดสอบสมมติฐาน โดยมีการทดสอบ t-test การทดสอบ Pearson’s chi-square การทดสอบ Berry-Rogghe’s z-score การทดสอบค่า likelihood ratio การใช้ค่า Mutual Information การทดสอบแบ

บอื่นๆ กาหาคำปรากฏร่วมมากกว่าสองคำ การหาคำปรากฏร่วมระหว่างสองภาษา และ12.5 สรุป

บทที่ 13 การใช้เว็บเป็นคลังข้อมูล เนื้อหาในบทนี้ผู้เขียนได้นำเสนอเกี่ยวกับข้อมูลที่ปรากฏในเวบไซต์ต่างๆเป็นจำนวนมาก อธิบายเชิงตอบคำถามซึ่งเป็นที่สนใจของคนจำนวนมากเกี่ยวกับข้อมูลในเว็บไซต์ทั้งหมดเป็นคลังข้อมูลภาษาด้วยหรือไม่ ถ้าไม่ใช่แล้วยังสามารถนำมาใช้ประโยชน์ได้หรือไม่และใช้อย่างไร เป็นต้น การอธิบายเชิงตั้งคำถามของผู้เขียนในบทนี้แบ่งออกเป็นหัวข้อหลักๆดังนี้ 13.1) บทนำ 13.2) ข้อมูลภาษาในเว็บ 13.3) การรวบรวมข้อมูลเว็บสร้างเป็นคลังข้อมูลภาษา 13.4) การใช้เว็บเป็นคลังข้อมูลภาษา 13.5) WEB-BASED คอนคอร์แด๊นซ์ ซึ่งประกอบไปด้วย โปรแกรม WebCorp และโปรแกรม KWiCFiner และ 13.6 สรุป

สิ่งที่เป็นจุดเด่นในเนื้อหาบทที่ 11-13 นั้นผู้เขียนได้อธิบายถึงโปรแกรมที่ใช้สำหรับการค้นหาคำที่ต้องการและจัดเรียงคำในเว็บต่างๆ ซึ่งไม่ได้เป็นเพียงแค่การยกตัวอย่างโปรแกรมนั้น แต่ผู้เขียนยังได้นำเสนอวิธีการใช้โปรแกรมและนำเสนอตัวอย่างการใช้โปรแกรมการศึกษาคลังข้อมูลอีกด้วย ซึ่งการนำเสนอตัวอย่างการศึกษาดังกล่าวนี้ก็มีหลากหลายทำให้ผู้อ่านนั้นสามารถเรียนรู้และทำความเข้าใจได้ง่ายมากยิ่งขึ้นด้วย

ความสำคัญและประโยชน์ของของหนังสือ

หนังสือเล่มนี้มีความสำคัญและมีประโยชน์ดังนี้

1) เป็นหนังสือ “ภาษาศาสตร์คลังข้อมูล: หลักการและการใช้” นับได้ว่าเป็นหนังสือวิชาการที่อธิบายเกี่ยวกับคลังข้อมูลภาษาและการศึกษาภาษาศาสตร์ในคลังข้อมูลเป็นภาษาไทยเล่มแรกๆที่มีการเผยแพร่ในประเทศไทย เนื่องจากโดยส่วนมากจะมีปรากฏเฉพาะหนังสือที่เป็นภาษาอังกฤษเท่านั้น หนังสือเล่มนี้ผู้เขียนใช้เวลารวบรวมข้อมูลและพัฒนามาจากรายงานการวิจัยเรื่อง แนวทางการพัฒนาคลังข้อมูล: กรณีศึกษาจากการสร้างคลังข้อมูลภาษาอังกฤษ ซึ่งผู้เขียนได้รับทุนสนับสนุนการวิจัยจากฝ่ายวิจัย คณะอักษรศาสตร์ในปี พ.ศ. 2541 หนังสือเล่มนี้ตีพิมพ์ครั้งแรกในปี พ.ศ. 2545 ต่อมามีการปรับปรุงแก้ไขเพื่อให้เกิดความสมบูรณ์และมีความทันสมัยอยู่เสมอ และในปี พ.ศ. 2553 จึงได้ดำเนินการจัดพิมพ์ครั้งที่ 2

2) ผู้เขียนร้อยเรียงภาษาเพื่ออธิบายเนื้อหาข้อมูลด้วยภาษาที่เรียบง่าย เข้าใจง่าย

ลื่นไหลต่อเนื่องและไม่สะดุด จึงทำให้หนังสือวิชาการซึ่งเป็นศาสตร์เฉพาะทางด้านภาษาศาสตร์ซึ่งค่อนข้างยากอยู่แล้ว แต่ผู้เขียนสามารถทำให้เป็นเรื่องง่ายและน่าติดตาม

3) หัวข้อแต่ละหัวข้อ เนื้อหาการอธิบายผู้เขียนจะยกตัวอย่างชื่อพจนานุกรมต่างๆ ขั้นตอนการจัดทำ โปรแกรมคอมพิวเตอร์ในการรวบรวม เรียบเรียงอย่างละเอียด โดยใช้คำและภาษาที่เข้าใจง่าย นอกเหนือจากนั้นขั้นตอนในแต่ละหัวข้อที่ผู้เขียนได้หยิบยกนำมาอธิบายนั้นจะมีภาพประกอบแสดงตัวอย่างกำกับไว้ด้วย วิธีการนี้ทำให้ผู้อ่านสามารถมองเห็นภาพและสามารถเข้าใจได้อย่างง่าย

4) ข้อมูล หนังสือ เอกสารต่างๆซึ่งผู้เขียนได้นำมาใช้สำหรับการเขียนอ้างอิงในหนังสือเล่มนี้มีจำนวนมากถึง 211 รายการ และรายการอ้างอิงทั้งหมดล้วนเป็นภาษาอังกฤษ จึงทำให้ข้อมูลในหนังสือนั้นมีความชัดเจนและน่าเชื่อถือเป็นอย่างมาก

5) หนังสือเล่มนี้เป็นหนังสือที่สามารถนำมาใช้ศึกษาการจัดพจนานุกรมได้ด้วยตัวเองเนื่องจากผู้เขียนได้นำเสนอเป็นขั้นตอนด้วยกระบวนการวิธีง่ายๆและสามารถทำได้จริง

6) มีเนื้อหาที่ทันสมัย ข้อมูล วิชาการ ตัวอย่างเหมาะสมและมีความสอดคล้องกับเหตุการณ์ในปัจจุบันซึ่งการเรียนการสอนภาษาหรือการศึกษาค้นคว้า งานวิจัยเกี่ยวกับภาษานั้นคลังข้อมูลภาษามีความสำคัญอย่างมาก

ข้อเสนอแนะสำหรับผู้อ่าน

หนังสือเล่มนี้ชี้ให้เห็นถึงให้บทบาทและความสำคัญของคลังข้อมูลภาษาในงานด้านต่างๆ ได้แก่ การวิจัยทางภาษาศาสตร์ การเรียนการสอนภาษา การแปลภาษา และการประมวลผลภาษา เป็นหนังสือที่มีความเหมาะสมสำหรับนักวิชาการ นักศึกษา และผู้มีความสนใจทั่วไปที่จะเรียนรู้ภาษาศาสตร์อีกแขนงหนึ่งซึ่งเกี่ยวข้องกับการรวบรวมข้อมูลภาษาเพื่อจัดทำคลังข้อมูลภาษาสำหรับใช้งานโดยใช้โปรแกรมทางคอมพิวเตอร์มาประยุกต์ใช้ร่วม ดังนั้นนักวิชาการ นักศึกษาและผู้มีความสนใจทั่วไปไม่ควรพลาดที่จะสรรหาเพื่อนำมาเก็บไว้ศึกษา ใช้เป็นหนังสืออ้างอิงในงานวิชาการและเป็นสมบัติส่วนตัว

เอกสารอ้างอิง

วิโรจน์ อรุณมานะกุล. (2553). *ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้.* (พิมพ์ครั้งที่ 2). กรุงเทพฯ : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.