บทที่ 3
การสร้างคลังข้อมูลภาษา

คลังข้อมูลภาษา (Corpus) เป็นสิ่งสำคัญยิ่งในการจัดทำประมวลศัพท์ เพราะนอกจากจะเป็นฐานข้อมูลการใช้ศัพท์จริงที่รวบรวมมาแล้ว ยังเป็นแหล่งข้อมูลที่ทำให้ทราบความหมายและการใช้งานของศัพท์ได้อีกด้วย การใช้คลังข้อมูลภาษาในการจัดทำประมวลศัพท์มีมาตั้งแต่ในอดีต แต่มาเริ่มได้รับความนิยมในปัจจุบัน ความสำคัญและความสะดวกที่คลังข้อมูลภาษามีต่อกระบวนการทำประมวลศัพท์ ทำให้ในปัจจุบันคลังข้อมูลภาษาได้กลายเป็นเครื่องมือที่สำคัญที่สุดอย่างหนึ่งในการจัดทำประมวลศัพท์ ซึ่งผู้จัดทำประมวลศัพท์ทุกคนควรให้ความสำคัญ

เนื้อหาในบทนี้จะเป็นการพูดถึงประวัติความเป็นมาโดยคร่าวๆของการใช้คลังข้อมูลภาษาในการสร้างประมวลศัพท์ ความสำคัญของคลังข้อมูลภาษาในการสร้างประมวลศัพท์ วิธีวิทยาในการสร้างคลังข้อมูลภาษา รวมถึงปัญหาและวิธีการแก้ไขที่ประสบในการดำเนินการดังกล่าวด้วย โดยคำนึงถึงกระบวนการที่ปฏิบัติจริงในการสร้างประมวลศัพท์เรื่องการเติมน้ำลงชั้นน้ำบาดาลเป็นหลัก

ความเป็นมาของการใช้คลังข้อมูลภาษาในการทำประมวลศัพท์

แม้ว่าคลังข้อมูลภาษาจะเป็นสิ่งสำคัญและจำเป็นดังที่กล่าวมาแล้ว แต่ก็ยังมีผู้ไม่เห็นด้วยกับการใช้คลังข้อมูลภาษาเป็นเครื่องมือในการสร้างประมวลศัพท์ Pearson (1998) ได้กล่าวไว้ว่า เดิมทีนักศัพทวิทยาไม่นิยมใช้คลังข้อมูลภาษาในการจัดทำประมวลศัพท์ สาเหตุที่เป็นเช่นนั้นเพราะในอดีตการสร้างคลังข้อมูลภาษาเฉพาะด้าน (Specialized Corpus) มีความยากลำบาก โดยเฉพาะในแง่ของการรวบรวมข้อมูลภาษาเป็นจำนวนมากๆ ต่อมาเมื่อมีการเก็บข้อมูลในรูปของอิเล็กทรอนิก (Electronic Text) และมีการนำเทคโนโลยีคอมพิวเตอร์มาช่วยในการรวบรวมข้อมูล (ดูรายละเอียดในบทที่ 7) จึงเป็นการง่ายขึ้นที่จะสร้างคลังข้อมูลภาษาเฉพาะด้านของตนเองขึ้นมา แต่ก็ยังมีอุปสรรคอีกด้านหนึ่งซึ่งเป็นอุปสรรคต่อการใช้คลังข้อมูลภาษาในกระบวนการสร้างประมวลศัพท์ นั่นคือความเชื่อของนักศัพทวิทยาบางกลุ่มที่ยังไม่เชื่อมั่นในประโยชน์ของการใช้คลังข้อมูลภาษา แต่ในขณะเดียวกันก็มีนักศัพทวิทยารุ่นใหม่ที่มองเห็นประโยชน์และความสำคัญของการใช้คลังข้อมูลภาษาในการทำงาน ดังนั้น จึงมีการแบ่งนักศัพทวิทยาออกเป็นสองกลุ่มด้วยกันคือ นักศัพทวิทยากลุ่มอนุรักษ์นิยม (Traditional Terminologists) ที่เชื่อมั่นว่าศัพท์ (Terms) นั้นเป็นสิ่งที่เป็นอิสระจากบริบท (Context) อย่างเด็ดขาด และข้อมูลที่ได้จากคลังข้อมูลภาษาจะไม่เกิดประโยชน์แต่อย่างใดต่อ

การจัดทำประมวลศัพท์ กับอีกกลุ่มคือ นักศัพทวิทยากลุ่มรุ่นใหม่ (Modern Terminologists) ที่เชื่อว่าบริบทสามารถบ่งชี้รายละเอียดการใช้ศัพท์และความหมายของศัพท์ได้ และสนับสนุนให้มีการจัดทำคลังข้อมูลภาษาเพื่อใช้ในกระบวนการศัพทวิทยา

เมื่อเวลาผ่านไป ความคิดของนักศัพทวิทยาในกลุ่มรุ่นใหม่ได้รับการพัฒนาและสนับสนุนขึ้นตามลำดับ ทำให้การใช้คลังข้อมูลภาษาในกระบวนการศัพทวิทยาเริ่มได้รับการยอมรับมากขึ้นเรื่อยๆ และมีผู้นิยมนำมาใช้ในการสร้างประมวลศัพท์มากขึ้น จวบจนปัจจุบัน การจัดทำประมวลศัพท์เกือบทั้งหมดไม่สามารถหลีกเลี่ยงการใช้คลังข้อมูลภาษาได้อีกต่อไป เพราะนอกจากจะช่วยเพิ่มความสะดวกสบายและรวดเร็วให้กับการทำงานในหลายๆ ขั้นตอนแล้ว ก็ยังมีการพิสูจน์แล้วว่าข้อมูลบริบทที่ได้จากคลังข้อมูลภาษาสามารถบอกรายละเอียดเกี่ยวกับศัพท์ได้มาก ทั้งในแง่ของนิยาม (Definition) ตัวอย่างการใช้ (Example) คำเหมือน (Synonym) หรือคำตรงข้าม (Antonym) เป็นต้น อย่างไรก็ตาม การนำคลังข้อมูลภาษามาใช้ในกระบวนการทำประมวลศัพท์ก็นับเป็นก้าวสำคัญในการพัฒนาศาสตร์ด้านศัพทวิทยาไปอีกขั้นหนึ่ง ประโยชน์และรายละเอียดการนำคลังข้อมูลภาษามาใช้ในการสร้างประมวลศัพท์จะได้กล่าวโดยละเอียดต่อไป

ความสำคัญของคลังข้อมูลภาษาในกระบวนการทำประมวลศัพท์

คลังข้อมูลภาษานับว่ามีความสำคัญยิ่งในกระบวนการจัดทำประมวลศัพท์ ขั้นตอนการสร้างคลังข้อมูลภาษาเป็นขั้นตอนที่ผู้จัดทำประมวลศัพท์ต้องใช้ความระมัดระวังและรอบคอบอย่างมาก เพราะผลการดำเนินการในขั้นตอนนี้จะมีผลกระทบต่อขั้นตอนต่อๆ ไปทั้งหมดของการสร้างประมวลศัพท์ และยังมีผลโดยตรงต่อคุณภาพของประมวลศัพท์ที่ได้ในขั้นตอนสุดท้าย ซึ่งความสำคัญของคลังข้อมูลภาษาก็ขึ้นอยู่กับปัจจัยหลายอย่างด้วยกัน

ความสำคัญประการแรกของคลังข้อมูลภาษาก็คือ ข้อมูลที่ผู้จัดทำคลังข้อมูลภาษาได้รวบรวมไว้นั้นเป็นข้อมูลภาษาของศาสตร์เฉพาะด้านที่ต้องการนำมาสร้างเป็นประมวลศัพท์ ดังนั้นข้อมูลที่อยู่ในคลังข้อมูลภาษาจึงเปรียบเสมือนแหล่งข้อมูลอย่างดีเกี่ยวกับศัพท์เฉพาะด้านในเรื่องนั้นๆ ซึ่งข้อมูลนี้เองจะเป็นสิ่งที่ผู้ทำประมวลศัพท์จะนำมาใช้ในการบันทึกข้อมูลเบื้องต้นก่อนที่จะนำไปรวบรวมไว้ในบันทึกข้อมูลศัพท์อีกชั้นหนึ่ง ข้อมูลที่ได้จากคลังข้อมูลภาษานี้มาจากการใช้งานจริงซึ่งเป็นข้อมูลประเภทที่ผู้ทำประมวลศัพท์ต้องการที่สุด และยังเป็นการสนองตอบต่อวัตถุประสงค์ของการทำวิจัย นั่นคือการรวบรวมศัพท์ในรูปแบบที่มีการใช้งานจริง

นอกจากความสำคัญในแง่ของข้อมูลศัพท์แล้ว คลังข้อมูลภาษายังเป็นปัจจัยที่กำหนดความหลากหลายของประมวลศัพท์ด้วย การกำหนดเกณฑ์ในการเลือกข้อมูลที่จะมาเก็บไว้ในคลังข้อมูลภาษาจะมีผลโดยตรงต่อประมวลศัพท์ที่ออกมา เช่น ถ้ากำหนดว่าข้อมูลในประมวลศัพท์จะเก็บเฉพาะข้อมูลจากตำราเรียนเท่านั้น ความหลากหลายของศัพท์อาจมีน้อยกว่าข้อมูลที่เก็บจากหลายๆ แหล่ง เช่น บทความทางวิชาการ เอกสารเทคนิค หรือ ข่าวตามหนังสือพิมพ์ เป็นต้น นอกจากความหลากหลายแล้ว ยังมีผลกระทบต่อระดับภาษาของศัพท์ที่ได้ จำนวนศัพท์ และความหนาแน่นของศัพท์ในข้อมูลนั้นๆ อีกด้วย

ข้อมูลจากคลังข้อมูลภาษาจะเป็นข้อมูลหลักในการทำบันทึกข้อมูลศัพท์ ไม่ว่าจะเป็นข้อมูลด้านไวยากรณ์ (Grammatical Category) ด้านตัวอย่างการใช้งาน ด้านข้อมูลเสริม เช่น คำเหมือน (Synonym) และคำตรงข้าม (Antonym) เป็นต้น หรือแม้แต่ความหมายของศัพท์เองก็ตาม ดังนั้นด้วยความสำคัญดังที่กล่าวมาทั้งหมดนี้ของคลังข้อมูลภาษา การจัดทำคลังข้อมูลภาษา หรือวิธีวิทยาในการจัดทำคลังข้อมูลภาษา จึงเป็นขั้นตอนที่ต้องอาศัยการวางแผนและกำหนดเป้าหมายไว้อย่างดีเพื่อให้เกิดอุปสรรคปัญหาน้อยที่สุดในการทำงาน และเพื่อให้ได้คลังข้อมูลภาษาที่มีคุณภาพเพื่อเป็นฐานที่แข็งแกร่งให้กับการจัดทำประมวลศัพท์ต่อไป

วิธีวิทยาในการสร้างคลังข้อมูลภาษา

วิธีวิทยาในการสร้างคลังข้อมูลภาษาแบ่งออกเป็นสองส่วนคือ ส่วนของการคัดเลือกข้อมูลเพื่อมาเก็บรวบรวม และส่วนของการจัดเก็บและจัดระบบข้อมูลเพื่อให้พร้อมใช้งาน

1. การรวบรวมข้อมูลเพื่อสร้างคลังข้อมูลภาษา

คลังข้อมูลภาษาเกิดจากการรวบรวมข้อมูล (Text) เกี่ยวกับศาสตร์เฉพาะด้านที่มีคุณสมบัติตามที่ผู้รวบรวมต้องการ มาไว้รวมกันเพื่อความสะดวกในการใช้งานด้านศัพทวิทยา ซึ่งเกณฑ์ในการเลือกข้อมูลหลักๆ ที่ต้องคำนึงถึงก็คือ ขนาดของคลังข้อมูลภาษา หัวข้อของข้อมูล และประเภทของตัวบท เกณฑ์นอกจากนั้นก็ขึ้นอยู่กับสถานการณ์การรวบรวมข้อมูล และวัตถุประสงค์ของประมวลศัพท์ตามแต่กรณีไป โดยทั่วไปการเลือกข้อมูลมาเก็บไว้ในคลังข้อมูลภาษามักจะใช้หลักเกณฑ์สำคัญสองประการด้วยกัน (Pearson, 1998: 52-54) ได้แก่

1) เกณฑ์ภายนอก (External Criteria) หรือบางตำราก็เรียกว่า Non-Linguistic Criteria หรือ Sociocultural Criteria ซึ่งเกณฑ์ชนิดนี้ประกอบด้วยองค์ประกอบดังนี้

- ประเภทของข้อมูล (Genre) เช่น เป็นข้อมูลจากตำราเรียน บทความหนังสือพิมพ์ หรือจากเอกสารเทคนิค เป็นต้น

- รูปแบบ (Mode) เช่น คำกล่าว (Speech) ข้อมูลเอกสาร (Written Text) ข้อมูลอิเล็กทรอนิกส์ (Electronic Text) เป็นต้น

- ผู้เขียน (Origin) เช่น ผู้เชี่ยวชาญ (Expert) นักเทคนิค (Technician) หรือ อาจารย์ เป็นต้น

- ผู้อ่านเป้าหมาย (Target Reader) เช่น กลุ่มบุคคลทั่วไป นักเทคนิค หรือ นักเรียน เป็นต้น

- จุดมุ่งหมาย (Aims of Text) เช่น ให้ข้อมูล (Informative) หารือ (Discussion) หรือ ให้คำแนะนำ (Recommendation) เป็นต้น

ซึ่งหลักเกณฑ์ภายนอกเหล่านี้ ผู้จัดทำประมวลศัพท์ต้องเป็นผู้ที่กำหนดขึ้นให้สอดคล้องกับกลุ่มเป้าหมาย วัตถุประสงค์ของประมวลศัพท์ และประเภทข้อมูลที่มีอยู่

2) เกณฑ์ภายใน (Internal Criteria) มีองค์ประกอบดังนี้

- หัวเรื่อง (Topic) ข้อมูลที่รวบรวมต้องอยู่ในหัวข้อเรื่องที่กำหนดไว้ เช่น ข้อมูลที่รวบรวมสร้างคลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาล ก็ต้องมีเนื้อหาอยู่ในเรื่องการเติมน้ำลงชั้นน้ำบาดาลเท่านั้น

- รูปแบบการใช้ภาษา (Style) เช่น ภาษาแบบเป็นทางการ (Formal) หรือ ภาษาแบบไม่เป็นทางการ (Informal) เป็นต้น

ในการจัดทำคลังข้อมูลภาษาสำหรับประมวลศัพท์เรื่องการเติมน้ำลงชั้นน้ำบาดาล ผู้ทำการวิจัยได้กำหนดเกณฑ์การเลือกข้อมูล (Text) โดยคำนึงถึงเกณฑ์พื้นฐานข้างต้น วัตถุประสงค์ของการทำประมวลศัพท์ และข้อจำกัดความเหมาะสมต่างๆ ในสถานการณ์การวิจัย และได้ข้อสรุปดังนี้

1) ขนาดของคลังข้อมูลภาษา (Size)

ข้อมูลเฉพาะด้านที่รวบรวมมาทั้งหมดกำหนดให้มีจำนวนประมาณ 100,000 คำ ซึ่งเป็นจำนวนที่คำนึงถึงความสำคัญสองด้าน คือ ด้านปริมาณข้อมูลที่จะหาได้ และคุณภาพของคลังข้อมูลภาษา

ในแง่ของปริมาณข้อมูล ศาสตร์ด้านการเติมน้ำลงชั้นน้ำบาดาลเป็นศาสตร์ทางด้านวิศวกรรมน้ำที่ยังใหม่อยู่มาก และมีข้อมูลเฉพาะด้านอยู่น้อยเมื่อเทียบกับศาสตร์เฉพาะด้านอื่นๆ นับเป็นอุปสรรคสำคัญในการสร้างคลังข้อมูลภาษาครั้งนี้ โดยเฉพาะอย่างยิ่งข้อมูลที่เป็นภาษาไทยยิ่งมีน้อยมาก เพราะประเทศไทยได้เริ่มหันมาสนใจศาสตร์การเติมน้ำลงชั้นน้ำบาดาลเพียงไม่กี่ปีมานี้เท่านั้น (โครงการเติมน้ำลงชั้นน้ำบาดาลจัดตั้งขึ้นในปี พ.ศ.2542) ดังนั้นจึงมีข้อมูลที่เป็นเอกสารน้อย นับเป็นสาเหตุหนึ่งที่ทำให้ไม่สามารถใช้กระบวนการทำวิจัยแบบ Multilingual Searches ได้ (ดูบทที่ 2) เพราะปริมาณข้อมูลของสองภาษาต่างกันมากเกินไป ปริมาณข้อมูลที่มีจำกัดนี้ทำให้ไม่สามารถสร้างคลังข้อมูลภาษาที่ใหญ่มากได้ แต่จำนวน 100,000 ก็ไม่ใช่จำนวนที่น้อยเกินไป เมื่อคำนึงถึงในแง่ของคุณภาพข้อมูลด้วย

สำหรับปัจจัยด้านคุณภาพของคลังข้อมูลภาษา ตามหลักศัพทวิทยาแล้วยิ่งคลังข้อมูลภาษาใหญ่เท่าไรก็ยิ่งดีเท่านั้น แต่ในกรณีที่คลังข้อมูลภาษามีขนาดไม่ใหญ่มากเช่นนี้ คุณภาพของข้อมูลจะเป็นสิ่งที่ชดเชยได้ ซึ่งเกณฑ์การเลือกข้อมูลที่ได้ตั้งไว้สำหรับคลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาลนี้ คาดว่าจะช่วยให้คลังข้อมูลภาษาที่ได้มีคุณภาพพอสมควร และชดเชยขนาดของคลังข้อมูลภาษาได้ ซึ่งในทฤษฎีเกี่ยวกับศัพทวิทยาบางทฤษฎี ก็ยังไม่สามารถกำหนดไว้อย่างแน่นอนตายตัวได้เลยว่า ขนาดของคลังข้อมูลภาษาควรจะมีทั้งหมดกี่คำจึงจะเป็นปริมาณที่เหมาะสมที่สุดในการจัดทำประมวลศัพท์ (Pearson, 1998) เพราะเมื่อพิจารณาดูแล้ว การกำหนดขนาดของข้อมูลภาษาควรจะพิจารณาเป็นรายกรณีไป ซึ่งในแต่ละกรณีก็อาจแตกต่างกันไปขึ้นอยู่กับองค์ประกอบและปัจจัยแวดล้อมในขณะนั้นๆ ด้วย

2) หัวข้อ (Topic)

หลักเกณฑ์นี้มีขึ้นเพื่อเน้นย้ำว่า ข้อมูลที่รวบรวมมาต้องเป็นข้อมูลในเรื่องการเติมน้ำลงชั้นน้ำบาดาลเท่านั้น วิธีตรวจสอบเพื่อความแน่ใจว่าเป็นข้อมูลในหัวข้อดังกล่าวจริงทำได้ 2 ขั้นตอน คือ ตรวจสอบเนื้อหาของข้อมูล แล้วจึงนำไปสอบถามจากผู้เชี่ยวชาญศาสตร์เฉพาะด้านอีกครั้ง

นอกจากการตรวจสอบว่าเป็นข้อมูลในเรื่องที่ต้องการหรือไม่แล้ว อีกสิ่งหนึ่งที่ต้องคำนึงถึงในส่วนของหัวข้อก็คือ ขอบเขตของหัวข้อ ซึ่งจะเป็นขอบเขตของข้อมูลในคลังข้อมูลภาษาไปในตัว เรื่องการเติมน้ำลงชั้นน้ำบาดาลแบ่งได้เป็นหลายมิติด้วยกัน เช่น ด้านการบริหารจัดการ ด้านวิธีปฏิบัติการ หรือด้านเทคนิควิศวกรรม เป็นต้น ซึ่งผู้ทำวิจัยต้องกำหนดไว้ว่าต้องการขอบเขตของข้อมูลเพียงใด ทั้งนี้เพื่อไม่ให้ข้อมูลที่ได้กระจัดกระจายขาดระบบ เมื่อพิจารณาดูจาก

กลุ่มเป้าหมายและวัตถุประสงค์ของการทำประมวลศัพท์ครั้งนี้แล้ว (ดูบทนำ) จึงพอสรุปได้ว่าขอบเขตของข้อมูลควรจะอยู่ในระดับไม่กว้างหรือแคบเกินไป วงศัพท์ควรครอบคลุมศัพท์ที่ใช้กันส่วนใหญ่ โดยอาจมีการลงรายละเอียดในด้านเทคนิคหรือวิศวกรรมศาสตร์ที่พบใช้กันบ่อยๆ ในศาสตร์ด้านนี้ด้วย แต่ก็เพียงในปริมาณและขอบเขตที่กลุ่มเป้าหมายที่เป็นคนทั่วไปที่ไม่ใช่นักวิชาการหรือผู้เชี่ยวชาญเฉพาะด้านอาจทำความเข้าใจได้ และอาจต้องพบเจอในการเข้ามาเกี่ยวข้องกับเรื่องการเติมน้ำลงชั้นน้ำบาดาลเท่านั้น ดังนั้น จึงสรุปได้ว่าข้อมูลที่จะรวบรวมไว้ในคลังข้อมูลภาษาเป็นข้อมูลทั่วๆไปที่ครอบคลุมสาระสำคัญทั้งหมดในเรื่องการเติมน้ำลงชั้นน้ำบาดาลเอาไว้ แต่จะหลีกเลี่ยงข้อมูลเฉพาะด้านที่ลึกซึ้งมากๆ และจะใช้กันเฉพาะในหมู่ผู้เชี่ยวชาญเฉพาะด้านจริงๆ เช่น เทคนิคการเจาะบ่อบาดาล หรือ ข้อมูลการทดสอบความเป็นกรดด่างของน้ำในบ่อสูบ เป็นต้น อย่างไรก็ตามศัพท์เทคนิคหรือศัพท์วิทยาศาสตร์ส่วนหนึ่งก็จะไปปรากฏอยู่ในคลังข้อมูลภาษาด้วย ทั้งนี้เพราะข้อมูลเหล่านั้นเป็นข้อมูลที่สำคัญในศาสตร์ด้านนี้หรือต้องมีการกล่าวอ้างถึงบ่อยๆ เช่น ศัพท์เกี่ยวกับส่วนประกอบบางอย่างของบ่อเติมน้ำ หรือกระบวนการเป่าล้างบ่อ เป็นต้น ซึ่งในกรณีเช่นนี้ก็ต้องรวบรวมไว้เพราะเป็นข้อมูลพื้นฐานที่ผู้ใช้งานกลุ่มเป้าหมายมีโอกาสจะได้พอเจอสูงในการทำงานหรือเอกสารที่เผยแพร่ทั่วไปในเรื่องการเติมน้ำลงชั้นน้ำบาดาล

3) รายละเอียดของข้อมูล

ในเกณฑ์การเลือกข้อมูล สิ่งที่จะขาดไม่ได้เลยคือการกำหนดรายละเอียดของข้อมูลที่ต้องการนอกเหนือจากการคำนึงถึงขนาดของคลังข้อมูลภาษาและหัวข้อของข้อมูล ในการคัดเลือกข้อมูลเรื่องการเติมน้ำลงชั้นน้ำบาดาลนี้ ได้กำหนดเกณฑ์ไว้ดังนี้

- *รูปแบบ (Mode)* การรวบรวมข้อมูลครั้งนี้ได้กำหนดให้เก็บรวบรวมเฉพาะข้อมูลจากเอกสาร (Written Texts) และข้อมูลอิเล็กทรอนิกส์ (Electronic Texts) เท่านั้น สาเหตุที่เป็นเช่นนั้นเพราะต้องการให้รูปแบบของศัพท์ที่ได้มีความเป็นทางการหรือกึ่งทางการเพื่อประโยชน์ในการใช้งานทางวิชาการเป็นหลัก และสาเหตุอีกประการหนึ่งก็เพราะข้อมูลในรูปแบบอื่นหาได้ยากกว่าและมีปริมาณน้อยกว่าด้วย อีกทั้งศัพท์ที่ได้จากบทสนทนาหรือคำกล่าวมักจะเป็นศัพท์ที่ไม่เป็นทางการหรือบางครั้งก็มีการตัดหรือย่อ ซึ่งทำให้ศัพท์ที่ได้ขาดมาตรฐานไป ส่วนสาเหตุที่เลือกใช้ข้อมูลจากเอกสารก็เพราะมีความน่าเชื่อถือสามารถค้นหาที่มาได้ อีกทั้งศัพท์ที่ได้ก็มีความเป็นทางการและมีมาตรฐานการใช้ที่ค่อนข้างแน่นอนกว่า สำหรับข้อมูลอิเล็กทรอนิกส์ ซึ่งในที่นี้เน้นข้อมูลที่ได้จากอินเตอร์เน็ตเป็นหลัก มีประโยชน์ในแง่ของข้อมูลที่อยู่ในรูปไฟล์คอมพิวเตอร์ทำให้ง่ายต่อการเก็บรวบรวม แต่ก็อาจมีปัญหาด้านความน่าเชื่อถือได้ ดังนั้นในการเก็บข้อมูลอิเล็กทรอนิกส์จึงต้องมีการตรวจ

สอบอย่างละเอียดรอบคอบว่ามีแหล่งที่มาแน่นอนหรือไม่และมีความน่าเชื่อถือมากน้อยเพียงใดก่อนที่จะนำมาเก็บรวบรวม

- *ประเภทของข้อมูล (Genre)* ประเภทของข้อมูลในคลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาล กำหนดให้มีสามประเภทด้วยกัน ได้แก่

ประเภทที่หนึ่ง ข้อมูลจากตำราเรียน (Texts from Text Books)

ประเภทที่สอง ข้อมูลจากเอกสารสำหรับผู้ปฏิบัติงาน (Technical Texts)

ประเภทที่สาม ข้อมูลจากเอกสารเผยแพร่ประชาสัมพันธ์ (Publication Texts)

สาเหตุที่รวบรวมข้อมูลสามประเภทนี้ไว้ก็เนื่องจากเป็นข้อมูลที่แสดงสถานการณ์สื่อสารที่ต่างกันไป ประเภทที่หนึ่งแสดงข้อมูลภาษาที่ใช้ในสถานการณ์สื่อสารระหว่างอาจารย์กับนักเรียน ประเภทที่สองระหว่างผู้เชี่ยวชาญและผู้ปฏิบัติงานหรือนักเทคนิค และประเภทที่สามระหว่างผู้เชี่ยวชาญหรือนักเทคนิคกับคนทั่วไป ซึ่งการรวบรวมข้อมูลที่แสดงสถานการณ์สื่อสารต่างกันเช่นนี้ไว้จะช่วยให้ได้ศัพท์ที่มีความหลากหลาย และยังจะได้ข้อมูลเปรียบเทียบด้วยว่าในแต่ละสถานการณ์สื่อสารมีการใช้ศัพท์ต่างกันอย่างไรและจะมีผลต่อประมวลศัพท์ที่ได้มาอย่างไรบ้าง

- *ความน่าเชื่อถือ (Origin)* ข้อมูลในคลังข้อมูลภาษาทั้งหมดต้องเป็นข้อมูลที่มีแหล่งที่มาแน่นอน และมีความน่าเชื่อถือทางวิชาการ ผู้เขียนหรือสถาบันที่ผลิตข้อมูลออกมาต้องเป็นที่ยอมรับในวงการ การตรวจสอบตรงนี้ทำได้โดยการตรวจสอบที่มาของข้อมูล และนำไปหารือขอความเห็นกับผู้เชี่ยวชาญ เพื่อให้ข้อมูลที่ได้มีคุณภาพที่เชื่อถือได้เหมาะจะใช้เป็นฐานข้อมูลสำหรับการสร้างประมวลศัพท์

- *ความครบถ้วน* เนื่องจากศาสตร์ด้านการเติมน้ำลงชั้นน้ำบาดาลมีประเทศที่เป็นผู้บุกเบิก คือ สหรัฐอเมริกา แต่ต่อมาศาสตร์นี้ได้รับการเผยแพร่และพัฒนาขึ้นมากในประเทศแถบยุโรป และประเทศออสเตรเลีย ปัจจุบันกลุ่มประเทศทั้งสามต่างก็ออกเอกสารและข้อมูลเกี่ยวกับศาสตร์ด้านน้ำบาดาลมาอย่างต่อเนื่อง แต่ละสำนักก็มีชุดศัพท์ของตนเอง จึงนับได้ว่าศาสตร์ด้านนี้มีสำนักใหญ่ๆ ที่มีอิทธิพลต่อการบัญญัติใช้ศัพท์เฉพาะด้านอยู่สามสำนัก และเพื่อให้ประมวลศัพท์ที่ได้มีความครบถ้วนในแง่ของศัพท์ที่มีการใช้จริง การรวบรวมข้อมูลสร้างคลังข้อมูลภาษาจึงกำหนดให้เก็บรวบรวมข้อมูลที่มีที่มาจากทั้งสามสำนัก ซึ่งถ้าทำได้ประมวลศัพท์ที่ได้ก็จะนำไปใช้งานกับเอกสารที่เขียนโดยสำนักทั้งสามได้หมด และมีประโยชน์กว้างขวางมากกว่าที่จะจำกัดการใช้งานอยู่เพียงชุดศัพท์ของสำนักเดียว

- *คุณภาพ* (Quality) ข้อมูลที่เลือกมาควรเลือกที่มีความหนาแน่นของศัพท์เฉพาะด้านสูง (With high density of term) มีบริบทที่เป็นประโยชน์ ข้อมูลมีความชัดเจนแน่นอนและชี้ให้เห็นนิยาม (Definition) ของศัพท์ได้

- *ความทันสมัย (Up-to-date)* คลังข้อมูลภาษาที่ดีต้องมีความทันสมัย ข้อมูลที่รวบรวมมาจึงต้องไม่เก่ามากจนเกินไป เพื่อให้ประมวลศัพท์ที่ได้ไม่ล้าหลังจนใช้ประโยชน์ไม่ได้ เกณฑ์ที่กำหนดไว้สำหรับคลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาลคือ ต้องเป็นข้อมูลที่เขียนหรือรวบรวมมาไม่เกิน 10 ปี ยกเว้นแต่ข้อมูลจากตำราเรียนที่อาจเขียนมานานกว่า 10 ปีได้แต่ไม่เกิน 20 ปี เพราะตำราเรียนจะแสดงข้อมูลพื้นฐานเกี่ยวกับเรื่องการเติมน้ำลงชั้นน้ำบาดาล ซึ่งมีโอกาสเปลี่ยนแปลงได้น้อยกว่าข้อมูลแบบอื่น

2. การเก็บข้อมูลภาษา

เมื่อได้ข้อมูล (Texts) ตามหลักเกณฑ์ที่กำหนดไว้แล้ว ก็มาถึงขั้นตอนของการนำข้อมูลมารวบรวมไว้ด้วยกันเพื่อจัดเก็บให้อยู่ในรูปที่ใช้งานได้ง่ายและสะดวกที่สุด ซึ่งในปัจจุบันก็สามารถทำได้โดยการเก็บข้อมูลทั้งหมดไว้ในรูปของไฟล์คอมพิวเตอร์ และใช้ซอฟต์แวร์ที่ออกแบบมาโดยเฉพาะเพื่อจัดระบบไฟล์ทั้งหลายให้พร้อมใช้งานในขั้นตอนต่อๆ ไปของการทำประมวลศัพท์

2.1 การเปลี่ยนข้อมูลให้อยู่ในรูปของไฟล์คอมพิวเตอร์

ข้อมูลที่รวบรวมมามีสองรูปแบบตามที่ได้กล่าวมาแล้ว คือ ข้อมูลที่เป็นเอกสาร กับข้อมูลอิเล็กทรอนิกส์ สำหรับข้อมูลอิเล็กทรอนิกส์ย่อมไม่เป็นปัญหาเพราะอยู่ในรูปไฟล์อยู่แล้ว แต่ข้อมูลที่เป็นเอกสารจำเป็นต้องนำมาแปรสภาพให้อยู่ในรูปของไฟล์คอมพิวเตอร์ เพื่อให้ประหยัดพื้นที่ในการจัดเก็บ และเพื่อให้ใช้งานกับซอฟต์แวร์จัดการข้อมูลได้ วิธีที่ทำได้ก็คือ การนำข้อมูลเอกสารดังกล่าวไปเข้าเครื่องสแกนเนอร์เพื่อเปลี่ยนข้อมูลให้อยู่ในรูปไฟล์อิเล็กทรอนิกส์ ซึ่งโปรแกรมที่ใช้ในการเปลี่ยนนี้อาจใช้ได้หลายโปรแกรม เช่น OmniPage หรือ Text Bridge เป็นต้น (ปัญหาในขั้นตอนนี้ดูได้จากหัวข้อ 'ปัญหาและทางแก้ไข' ตอนท้ายบท)

2.2) การนำซอฟต์แวร์มาใช้ในการจัดระบบข้อมูล

ข้อมูลที่อยู่ในรูปของไฟล์คอมพิวเตอร์แล้วจะนำมาใช้ประโยชน์ในฐานะของคลังข้อมูลภาษาได้ก็โดยการนำซอฟต์แวร์ที่เหมาะสมมาใช้ในการจัดการข้อมูล เพื่อให้เกิดความสะดวกรวดเร็วในกระบวนการต่อๆ ไปของการทำประมวลศัพท์ ซอฟต์แวร์ดังกล่าวได้แก่ซอฟต์แวร์ใน

ตระกูล Concordance เช่น Win Concordance* ซึ่งเป็นโปรแกรมที่ใช้กันแพร่หลาย และเป็นโปรแกรมที่นำมาใช้งานในการทำวิจัยครั้งนี้ด้วย

เมื่อผ่านขั้นตอนการคัดเลือกข้อมูล รวมถึงการเก็บรวบรวมและจัดระบบข้อมูลแล้ว คลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาลก็เสร็จสมบูรณ์ พร้อมสำหรับการใช้งานในขั้นตอนต่อไป (รายละเอียดรายการอ้างอิงของข้อมูลในคลังข้อมูลภาษาเรื่องการเติมน้ำลงชั้นน้ำบาดาล ดูได้จากภาคผนวก ก)

ปัญหาในเรื่องการสร้างคลังข้อมูลภาษาและทางแก้ไข

แม้จะดูเหมือนว่าในขั้นตอนการสร้างคลังข้อมูลภาษานี้จะไม่มีความซับซ้อนมากนัก แต่เมื่อได้ลงมือทำวิจัยจริง ก็ประสบปัญหาในทางทฤษฎีและปฏิบัติไม่น้อย ซึ่งผู้ทำวิจัยก็ได้พยายามหาทางแก้ไข โดยมีรายละเอียดปัญหาและทางแก้ไขดังต่อไปนี้

ปัญหาในขั้นตอนการกำหนดเกณฑ์คัดเลือกข้อมูล

- *การค้นข้อมูล* การค้นข้อมูลเกี่ยวกับศาสตร์ใหม่อย่างการเติมน้ำลงชั้นน้ำบาดาลพบข้อมูลไม่มากอย่างศาสตร์เรื่องอื่นๆ วิธีแก้ไขคือ กระจายการค้นออกไปหลายๆ ด้าน เช่น ค้นจากฐานข้อมูลห้องสมุด ฐานข้อมูลของมหาวิทยาลัยทั้งในและนอกประเทศ รวมถึงข้อมูลทางอินเตอร์เน็ตด้วย ซึ่งการปฏิบัติเช่นนี้ทำให้สามารถรวบรวมข้อมูลมาได้จนครบตามเกณฑ์ที่กำหนดไว้

- *ข้อมูลที่ได้ขาดความหลากหลาย* จากการสำรวจข้อมูลเบื้องต้นเรื่องการเติมน้ำลงชั้นน้ำบาดาล พบว่าประเภทข้อมูลที่พบมากที่สุดคือ ตำราเรียน รองลงมาคือข้อมูลเทคนิค แต่ส่วนที่น้อยมากคือข้อมูลประชาสัมพันธ์ การจะสร้างคลังข้อมูลภาษาให้มีความสมดุลย์ได้ก็จำเป็นต้องหาข้อมูลด้านที่น้อยให้มีปริมาณใกล้เคียงกับอีกสองประเภท ทางแก้ที่ค้นพบก็คืออาศัยการค้นข้อมูลทางอินเตอร์เน็ต ซึ่งจะมีข้อมูลที่เป็นข่าวหรือบทความสั้นๆ ซึ่งจัดเป็นข้อมูลประชาสัมพันธ์ได้ เพราะเป็นสถานการณ์สื่อสารระหว่างผู้ที่มีความรู้เรื่องศาสตร์เฉพาะด้านกับบุคคลทั่วไปที่ไม่มีความรู้มาก่อน

---

* WinConcord เป็นโปรแกรม Concordancer for Window ที่พัฒนาขึ้นมาโดย Zdenek Martinek จาก the University of West Bohemia, Pilsen, Czech Republic ด้วยความร่วมมือกับ Les Siegrist, จาก the Technische Universität Darmstadt, Germany โปรแกรมนี้สามารถดาวน์โหลดได้จาก http://www.ifs.tu-darmstadt.de/sprachlit/wconcord.htm

- *ขนาดของคลังข้อมูลภาษา* ตามทฤษฎีการจัดทำประมวลศัพท์ส่วนใหญ่ มักจะแนะนำให้สร้างคลังข้อมูลที่มีจำนวนคำอย่างน้อย 1,000,000คำ (Pearson, 1998) ซึ่งถ้ามองในแง่นี้ คลังข้อมูลภาษาที่สร้างขึ้นในการวิจัยครั้งนี้ก็มีขนาดน้อยกว่าหลายเท่าตัว (100,000 คำ) แต่จากตัวอย่างการทำประมวลศัพท์ในต่างประเทศ มีอยู่หลายกรณีที่จำนวนศัพท์มีน้อย (50,000 – 100,000 คำ) แต่ก็สามารถผลิตผลงานที่มีคุณภาพได้ ดังนั้น ในความเห็นของผู้ทำวิจัยแล้ว ขนาดของคลังข้อมูลน่าจะขึ้นอยู่กับจุดประสงค์การใช้งานเป็นหลัก ในกรณีศึกษานี้ คลังข้อมูลภาษาขนาด 100,000 คำ มีความเหมาะสมแล้ว เพราะข้อมูลมีไม่มากนัก ข้อมูลที่มีอยู่มีคุณภาพสูง (มีความถี่ศัพท์มาก ข้อมูลแน่นอนเชื่อถือได้) และประมวลศัพท์เป้าหมายที่ตั้งไว้ก็อยู่ในขนาดกลางมีจำนวนศัพท์ไม่มากนัก (ประมาณ 100 ศัพท์) ดังนั้นแม้ว่าขนาดของคลังข้อมูลภาษาจะน้อยกว่าเกณฑ์มาตรฐานอยู่มาก แต่ก็มีองค์ประกอบและเหตุผลที่ชดเชยกันได้ดังที่กล่าวมาแล้ว

- *ที่มาของข้อมูลไม่แน่ชัด* ปัญหานี้เกิดกับข้อมูลส่วนใหญ่ที่ได้จากอินเตอร์เน็ต ข้อมูลเช่นนี้ส่วนใหญ่หาได้ง่ายแต่มักไม่สามารถยืนยันที่มาอย่างชัดเจนได้ ทำให้ข้อมูลจำนวนมากขาดคุณสมบัติด้านความน่าเชื่อถือซึ่งเป็นเกณฑ์หนึ่งในการคัดเลือกข้อมูล ทางออกสำหรับปัญหานี้คือต้องตัดข้อมูลที่ไม่มีแหล่งที่มาชัดเจนออกไป แม้จะน่าเสียดายแต่ต้องคำนึงถึงคุณภาพของข้อมูลเป็นหลักสำคัญ การแก้ปัญหานี้อาจทำให้ข้อมูลที่ได้มีน้อยลง แต่ก็จะช่วยป้องกันปัญหาใหญ่ๆ ที่จะเกิดตามมา เช่นปัญหาการอ้างอิงถึงแหล่งที่มาของข้อมูล ซึ่งอาจมีผลกระทบต่อคุณภาพและความน่าเชื่อถือของประมวลศัพท์ด้วย

ปัญหาในขั้นตอนการเก็บและจัดระบบข้อมูล

-ปัญหาค่าใช้จ่ายในการสแกนข้อมูลเอกสารให้อยู่ในรูปไฟล์คอมพิวเตอร์ ขั้นตอนนี้มีค่าใช้จ่ายมากในส่วนของการที่ต้องสแกนเอกสารจำนวนมาก ซึ่งถ้าไปจ้างร้านที่รับงานประเภทนี้ส่วนใหญ่จะคิดราคาประมาณแผ่นละ 10 บาท ผู้ทำวิจัยแก้ปัญหานี้ด้วยการใช้เครื่องสแกนเนอร์ส่วนตัว เป็นการลงทุนครั้งเดียว แต่ช่วยให้เกิดความสะดวกสบายในการทำงานขึ้นอีกมาก และเมื่อคำนวณถึงค่าใช้จ่ายในการสแกนและค่าเดินทางแล้ว การแก้ปัญหาเช่นนี้อาจจะคุ้มค่ากว่าในแง่ของค่าใช้จ่ายโดยรวม

- *ปัญหาซอฟต์แวร์ที่ใช้ในการเปลี่ยนข้อมูลเอกสารเป็นไฟล์คอมพิวเตอร์* ซอฟต์แวร์ที่ใช้เพื่อการนี้มีอยู่จำนวนหนึ่ง การเลือกใช้มีความสำคัญมากกับคุณภาพของข้อมูลที่จะได้มา ปัญหาของผู้ทำวิจัยเกิดจากการเลือกใช้โปรแกรม Text Bridge 6.0 ซึ่งแถมมากับเครื่องสแกนเนอร์ใน

การทำงานกับข้อมูลชุดแรก และประสบปัญหาทันที เนื่องจากคุณภาพของข้อมูลที่ออกมาค่อนข้างต่ำ ข้อมูลผิดพลาดไปจากต้นฉบับค่อนข้างมาก ทำให้ต้องเสียเวลามาปรับเปลี่ยนใหม่เป็นรายคำรายประโยค ทำให้เสียเวลาและอาจเกิดความผิดพลาดเนื่องจากการหลุดรอดสายตาไปได้โดยง่าย วิธีแก้ปัญหาคือ เปลี่ยนซอฟต์แวร์ที่ใช้ใหม่ ผู้ทำวิจัยเปลี่ยนจาก Text Bridge 6.0 มาใช้ OmniPage 8.0 และก็แก้ปัญหาข้างต้นไปได้มาก แม้ว่าโปรแกรมหลังนี้จะไม่เกิดความผิดพลาดมากเช่นโปรแกรมแรก แต่ก็มีจุดที่ผิดพลาดอยู่บ้าง ซึ่งต้องอาศัยการตรวจสอบ อย่างไรก็ตาม ถือว่าสามารถแก้ปัญหาส่วนใหญ่ไปได้แล้ว เพราะการสแกนก็ย่อมไม่สามารถคาดหวังให้ข้อมูลออกมาถูกต้อง 100 % ได้ การเลือกใช้ซอฟต์แวร์ที่มีคุณภาพสูงจะช่วยผ่อนภาระในส่วนนี้ไปได้มาก