โครงการเผยแพร่ผลงานวิชาการ
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ลำดับที่ 80

ชุด ภาษาศาสตร์เพื่อการงาน
ลำดับที่ 1

ภาษาศาสตร์คลังข้อมูล : หลักการและการใช้

วิโรจน์ อรุณมานะกุล

(พิมพ์ครั้งที่ 2)

คำนิยม

หนังสือ **ภาษาศาสตร์คลังข้อมูล** เล่มนี้ เป็นหนังสือที่มีคุณค่าและเหมาะสมแก่กาลเวลาเป็นอย่างยิ่ง ด้วยเหตุผลที่ว่าทั้งการศึกษาภาษาและการศึกษาด้านทฤษฎีภาษาศาสตร์ได้พัฒนามาถึงวาระที่เราไม่อาจจะอ้างอิงความถูกต้องของข้อสรุปหรือทฤษฎีที่นำเสนอ จากความรู้สึกหรือความสำนึกรู้ในฐานะเจ้าของภาษาเพียงอย่างเดียวได้อีกต่อไปแล้ว แต่จะต้องพยายามใช้ความเป็นจริงและข้อเท็จจริงของภาษามาเป็นหลักฐานสนับสนุนด้วย เครื่องมือหนึ่งที่จะช่วยให้เราเข้าถึงความเป็นจริงของภาษาได้ใกล้เคียงที่สุดก็คือคลังข้อมูลภาษา (Language corpus) ซึ่งรวบรวมข้อมูลภาษาที่ได้มีผู้ใช้ทั้งในการพูดและการเขียนมาเก็บรวบรวมไว้อย่างเป็นระบบเพื่อความสะดวกในการค้นหาข้อเท็จจริงที่จะนำไปพิสูจน์สมมุติฐาน ข้อสรุป หรือทฤษฎีของเรา ไม่ว่าเราจะทำงานด้านการศึกษาภาษามนุษย์ในภาพรวม ภาษาแต่ละภาษา ปรากฏการณ์ภาษา การแปลทั้งด้วยมือมนุษย์และด้วยเครื่อง การทำเครื่องมือช่วยในการสื่อภาษา เช่น พจนานุกรม และ อรรถาภิธาน ตลอดจนถึงการสร้างความสามารถทางด้านโปรแกรมที่ทำให้คอมพิวเตอร์สามารถประมวลผลภาษาธรรมชาติของมนุษย์ได้

วิวัฒนาการการสร้างและการใช้คลังข้อมูลภาษาดังกล่าวนี้ คงจะเกิดขึ้นไม่ได้หากไม่มีการนำวิวัฒนาการทางคอมพิวเตอร์มาผนวกเข้ากับวิวัฒนาการทางด้านภาษาศาสตร์ จนแตกเป็นแขนงขึ้นในสาขาภาษาศาสตร์ ที่เรียกขานกันว่า **ภาษาศาสตร์คอมพิวเตอร์** (Computational linguistics) และในสาขาวิทยาการคอมพิวเตอร์ ที่เรียกกันโดยรวมว่า **การประมวลภาษาธรรมชาติ** (Natural language processing) การผนวกกันของทั้งสองสาขาวิชาการนี้เป็นเรื่องที่ค่อนข้างยากลำบากมาแต่ต้น เนื่องจากมีนักวิชาการไม่

มากคนนักที่จะเชี่ยวชาญทั้งด้านภาษาศาสตร์และด้านคอมพิวเตอร์ อาจารย์ ดร.วิโรจน์ อรุณมานะกุล เป็นหนึ่งในผู้ที่มีความรู้อย่างลึกซึ้งในทั้งสองสาขาวิชาการ เนื่องด้วยเริ่มต้นมาจากพื้นฐานของนักคอมพิวเตอร์ในระดับปริญญาบัณฑิต และมาแตกแขนงเป็นวิชาภาษาศาสตร์ในระดับปริญญามหาบัณฑิต และต่อยอดอย่างสมบูรณ์ด้วยปริญญาดุษฎีบัณฑิตทางด้านภาษาศาสตร์คอมพิวเตอร์

ความเป็นนักวิชาการที่เชี่ยวชาญทั้งสาขาภาษาศาสตร์และสาขาคอมพิวเตอร์เพียงอย่างเดียว คงไม่เพียงพอที่จะทำให้หนังสือ **ภาษาศาสตร์คลังข้อมูล** นี้เกิดประโยชน์แก่นิสิตนักศึกษาและผู้สนใจได้อย่างเต็มที่ หากขาดเสียซึ่งความเป็นครูของผู้เขียน อาจารย์ ดร.วิโรจน์ อรุณมานะกุล เป็นครูที่ดีมาโดยตลอด นอกจากภูมิความรู้จะแน่นแล้ว ยังสอนหนังสือจากมุมของลูกศิษย์ผู้ยังไม่รู้มากนัก มิใช่จากมุมของความเป็นผู้เชี่ยวชาญของตน หนังสือเล่มนี้เขียนอย่างครูเขียนจึงอ่านง่ายและไม่ทำให้คนหน้าใหม่ที่อยากรู้หรือเพิ่งเริ่มศึกษาเรื่องนี้เกิดความรู้สึกระย่อเสียก่อน ส่วนผู้ที่มีความรู้อยู่บ้างแล้วไม่ว่าจะเป็นนักภาษา นักภาษาศาสตร์ หรือนักคอมพิวเตอร์ ก็สามารถจะได้ประโยชน์จากความสมบูรณ์ของเนื้อหาและแหล่งอ้างอิง ที่จะนำไปใช้ในการพัฒนาด้วยตนเองต่อไป

ขอแสดงความยินดีและความขอบคุณอาจารย์ ดร.วิโรจน์ อรุณมานะกุล ที่ได้เขียนหนังสือเล่มนี้ขึ้นมา เพื่อให้เป็นประโยชน์แก่ทุกคนที่สนใจเรื่องการศึกษาภาษาด้วยข้อมูลและหลักฐานอ้างอิงที่เป็นข้อเท็จจริง ไม่ว่าจะมาจากภูมิหลังอย่างใดและมีวัตถุประสงค์อย่างใดในการศึกษาภาษาของตน

เพียรศิริ วงศ์วิภานนท์ (Ph.D.)

คำนำ (ฉบับพิมพ์ครั้งที่ ๒)

หนังสือภาษาศาสตร์คลังข้อมูลตีพิมพ์ครั้งแรกในปีพ.ศ.2545 ได้ใช้ประกอบการสอนในวิชาที่เกี่ยวข้อง เช่น ภาษาศาสตร์คลังข้อมูล ทรัพยากรการแปล หลังจากที่หนังสือจำหน่ายหมดไปเมื่อ 2-3 ปีก่อน ผู้เขียนไม่ได้พิมพ์ใหม่ทันทีด้วยเห็นว่า ในระยะเวลาที่ผ่านมาได้มีประเด็นการศึกษาที่เกี่ยวข้องกับคลังข้อมูลภาษษเพิ่มเติมมากขึ้น อีกทั้งผู้เขียนตั้งใจจะเพิ่มในส่วนของการอธิบายและแสดงให้เห็นถึงการใช้โปรแกรมเครื่องมือต่างๆ ที่เกี่ยวข้องโดยละเอียดมากขึ้น จึงเป็นที่มาของฉบับปรับปรุงใหม่ที่ได้เพิ่มจำนวนหน้าจากเดิม 250 หน้าเป็นกว่า 380 หน้า

ด้วยเหตุนี้ หนังสือฉบับนี้จึงเปลี่ยนชื่อเรื่องใหม่เป็น *ภาษาศาสตร์คลังข้อมูล: หลักการและการใช้* เพราะตั้งใจจะนำเสนอทั้งสองแนวทางคือเรื่องของหลักการและเรื่องวิธีการใช้ แนวทางแรกเป็นเรื่องหลักการความคิดต่างๆ ที่เกี่ยวข้องกับคลังข้อมูลภาษาและการใช้ประโยชน์จากคลังข้อมูลซึ่งเป็นเนื้อหาหลักของฉบับที่พิมพ์ครั้งแรก โดยบทที่กล่าวถึงเรื่องประโยชน์ของการใช้คลังข้อมูลภาษาเดิมที่เป็นเรื่องการเรียนการสอนภาษา การวิจัยภาษา การแปลและการประมวลผลภาษา ได้ถูกแยกออกมาเป็นแต่ละบทในฉบับนี้ให้ชัดเจนขึ้นและได้เพิ่มประโยชน์ในการทำพจนานุกรม ตลอดจนกล่าวถึงคลังข้อมูลภาษาผู้เรียนและการใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียนเพิ่มขึ้นมาอีกหนึ่งบท บทนี้เป็นผลที่ได้มาจากส่วนหนึ่งของงานวิจัยที่ผู้เขียนทำโครงการวิจัยเรื่องคลังข้อมูลภาษาผู้เรียนภายใต้การสนับสนุนทุนวิจัยของสภาวิจัยแห่งชาติในระหว่างปี 2551-2553 ส่วนบทที่ว่าด้วยการกำกับข้อมูลภาษา ได้ปรับให้เน้นที่มาตรฐาน XML

เป็นหลักมากขึ้นแทนที่จะนำเสนอเฉพาะ SGML อย่างเดียว และในบทสุดท้ายได้ยกประเด็นที่ถกเถียงกันเรื่องของการใช้ประโยชน์จากเว็บว่าจะมาแทนที่คลังข้อมูลภาษาได้หรือไม่ และมีความเหมือนหรือความต่างกันอย่างไรระหว่างข้อมูลภาษาบนเว็บกับข้อมูลภาษาในคลังข้อมูล

ส่วนแนวทางที่สองที่เน้นเรื่องการใช้ ได้ปรับปรุงบทที่ว่าด้วยโปรแกรมคอนคอร์แด้นซ์โดยแสดงให้เห็นถึงขั้นตอนและตัวอย่างการใช้โปรแกรมไปด้วย และได้เพิ่มบทที่อธิบายเรื่องการนำข้อมูลภาษาเข้าในคอมพิวเตอร์เพื่อให้ผู้อ่านมีความเข้าใจที่ชัดเจนขึ้นว่า เมื่อนำข้อมูลภาษาเข้าไปในคอมพิวเตอร์แล้วข้อมูลภาษาจะถูกจัดเก็บอย่างไร เกี่ยวข้องกับเรื่องรหัสอักขระต่างๆอย่างไร จะนำข้อมูลเข้าได้ด้วยวิธีใดบ้าง และจะแปลงไฟล์ข้อมูลต่างๆที่มีอยู่ได้อย่างไรหากต้องการสร้างคลังข้อมูลภาษามาใช้เอง ตลอดจนได้แนะนำโปรแกรมที่ช่วยในการดาวน์โหลดข้อมูลจำนวนมากจากเว็บเพื่อความสะดวกในการรวบรวมข้อมูลภาษามาใช้ในงานวิจัย

เดิมทีผู้เขียนตั้งใจจะรอปรับแก้ไขหนังสือเล่มนี้ต่อไปอีกระยะด้วยเห็นว่ายังมีเนื้อหาใหม่ๆ ทางภาษาศาสตร์คลังข้อมูลเกิดขึ้นอยู่เรื่อยๆ แต่เนื่องจากหนังสือขาดตลาดมาเป็นเวลาพอสมควร จึงคิดว่าควรจะนำเสนอเท่าที่รวบรวมได้ใหม่นี้เสียก่อน หากมีเรื่องอื่นๆที่น่าสนใจพอก็อาจเขียนเป็นบทความหรือเป็นหนังสือต่างหากอีกเล่ม ท้ายที่สุดนี้ ผู้เขียนขอขอบคุณโครงการในแผนพัฒนาวิชาการ จุฬาลงกรณ์มหาวิทยาลัย (โครงการ จุฬาฯ 100 ปี) ที่ได้ให้ทุนสนันสนุนการจัดทำต้นฉบับหนังสือเล่มนี้ และโครงการเผยแพร่ผลงานวิชาการ ฝ่ายวิจัย คณะอักษรศาสตร์ที่ได้จัดพิมพ์และเผยแพร่หนังสือเล่มนี้

วิโรจน์ อรุณมานะกุล

21 มิถุนายน 2553

คำนำ (ฉบับพิมพ์ครั้งที่ ๑)

หนังสือภาษาศาสตร์คลังข้อมูลนี้ เขียนขึ้นเพื่อให้ความรู้พื้นฐานต่างๆ ที่เกี่ยวข้องกับคลังข้อมูลภาษา เพื่อให้ผู้อ่านได้เข้าใจและเห็นความสำคัญของการนำคลังข้อมูลภาษาไปใช้ประโยชน์ในการศึกษาด้านต่างๆ หนังสือเล่มนี้มีจุดเริ่มต้นมาจากรายงานการวิจัยเรื่อง “แนวทางการพัฒนาคลังข้อมูลภาษา: กรณีศึกษาจากการสร้างคลังข้อมูลภาษาอังกฤษ” ซึ่งผู้เขียนได้รับทุนสนันสนุนการวิจัยจากฝ่ายวิจัย คณะอักษรศาสตร์ ในปีพ.ศ. 2541 โครงการวิจัยนี้เป็นโครงการที่ศึกษากลวิธีที่ใช้ในการสร้างคลังข้อมูลภาษา โดยการสำรวจคลังข้อมูลภาษาจากแหล่งต่างๆ เพื่อวิเคราะห์ถึงลักษณะความแตกต่างรวมทั้งปัจจัยที่กำหนดลักษณะการจัดวางโครงสร้างคลังข้อมูลแต่ละรูปแบบ และสำรวจโปรแกรมซอฟท์แวร์ที่ใช้ร่วมกับคลังข้อมูลภาษาเหล่านั้น พร้อมทั้งทดลองสร้างคลังข้อมูลภาษา โดยคัดเลือกเอกสารอิเล็กทรอนิกส์ภาษาอังกฤษที่สามารถสืบค้นได้ทางอินเตอร์เน็ตและไม่มีข้อจำกัดเรื่องลิขสิทธิ์ เพื่อนำมารวบรวมเป็นคลังข้อมูลภาษาอังกฤษเพื่อใช้ประโยชน์ด้านการเรียนการสอนและงานวิจัยในคณะอักษรศาสตร์

หนังสือเล่มนี้ปรับปรุงขึ้นจากตอนที่หนึ่งของรายงานวิจัยดังกล่าว ซึ่งเป็นตอนที่เขียนขึ้นเพื่อให้พื้นฐานความรู้ที่จำเป็นทางด้านภาษาศาสตร์คลังข้อมูล ภาษาศาสตร์คลังข้อมูลเป็นวิชาที่ได้รับความสนใจอย่างมากในช่วงทศวรรษที่ 1990 นี้ โดยเน้นที่การศึกษาภาษาในแง่มุมต่างๆ โดยใช้ประโยชน์จากการรวบรวมข้อมูลภาษาจำนวนมาก คลังข้อมูลภาษาที่รวบรวมได้นี้สามารถใช้ประโยชน์ได้ทั้งในแง่ของการเรียนการสอนภาษา การวิจัยภาษา การแปล การจัดทำพจนานุกรม และการประมวลผลภาษาธรรมชาติ

บทที่ 1 เป็นบทแนะนำให้รู้ว่าคลังข้อมูลภาษาคืออะไร เพราะเหตุใดจึงได้รับความสนใจจากนักภาษาศาสตร์ และนักคอมพิวเตอร์เป็นอย่างมาก ซึ่งจะขยายความต่อในบทที่ 2 โดยจะกล่าวถึงประโยชน์ของคลังข้อมูลที่มีต่องานด้านต่างๆ สี่ด้าน ได้แก่ งานด้านการเรียนการสอนภาษา งานด้านการศึกษาวิจัยทางภาษาศาสตร์ งานด้านการแปล และงานทางด้านการประมวลผลภาษาธรรมชาติ เมื่อผู้อ่านเห็นภาพว่าคลังข้อมูลภาษาคืออะไรและมีประโยชน์อย่างไรแล้ว จึงจะได้กล่าวถึงการออกแบบและจัดสร้างคลังข้อมูลภาษาในบทที่ 3 โดยจะกล่าวถึงปัจจัยต่างๆที่ควรคำนึงในการออกแบบคลังข้อมูล ส่วนในบทที่ 4 เป็นเรื่องของการกำกับข้อมูลและมาตรฐานของการกำกับข้อมูล โดยจะกล่าวถึง SGML เป็นหลักเนื่องจากเป็นมาตรฐานที่ได้รับการยอมรับในปัจจุบันให้เป็นมาตรฐานของการกำกับข้อมูลในคลังข้อมูลภาษา แม้ว่าในปัจจุบัน จะมีผู้กล่าวถึงการใช้ XML ในการกำกับข้อมูลต่างๆ แต่เนื่องจาก SGML เป็นต้นกำเนิดของมาตรฐาน และ XML ก็ถือเป็น SGML ที่ถูกทำให้ง่ายขึ้น ผู้ที่เข้าใจ SGML แล้วก็จะสามารถศึกษา XML ได้โดยง่าย ในที่นี้ จึงได้นำเสนอเรื่องของ SGML เป็นสำคัญ

เนื่องจากการที่จะใช้ประโยชน์จากคลังข้อมูลภาษาให้ได้เต็มที่นั้น จำเป็นต้องอาศัยโปรแกรมหรือซอฟท์แวร์เพื่อช่วยในการค้น วิเคราะห์ หรือจัดสร้างคลังข้อมูล ในบทที่ 5 จึงกล่าวถึงโปรแกรมคอนคอร์แดนซ์ซึ่งเป็นโปรแกรมพื้นฐานสำหรับใช้ร่วมกับคลังข้อมูลภาษา เพื่อช่วยในการค้นหาคำและบริบทที่ปรากฏร่วมกับคำนั้นๆ โดยจะกล่าวถึงลักษณะทั่วไปของโปรแกรมประเภทนี้ และนำเสนอลักษณะเฉพาะของโปรแกรมคอนคอร์แดนซ์จำนวนหนึ่ง ส่วนบทที่ 6 จะกล่าวถึงวิธีการทางสถิติที่สามารถนำมาประยุกต์ใช้เพื่อช่วยหาคำที่มักปรากฏใช้ร่วมกัน ซึ่งเป็นการใช้ประโยชน์จากคลังข้อมูลภาษาในอีกลักษณะหนึ่ง

สำหรับส่วนที่สองและส่วนที่สามของงานวิจัยดังกล่าว ผู้เขียนไม่ได้นำมาเสนอในหนังสือเล่มนี้ เนื่องจากเป็นการรวบรวมแหล่งทรัพยากรที่เกี่ยวข้องกับคลังข้อมูลภาษาอังกฤษ และการจัดสร้าง CU English Text Archive ตามลำดับ ซึ่งไม่ได้เป็นส่วนที่ให้ความรู้พื้นฐานเกี่ยวกับศาสตร์ด้านนี้โดยตรง ผู้สนใจสามารถดาว์นโหลดรายงานวิจัยในส่วนนี้ได้จากอินเตอร์เน็ตที่

http://www.arts.chula.ac.th/~ling/paper/corp2.zip

http://www.arts.chula.ac.th/~ling/paper/corp3.zip

และสามารถใช้คลังข้อมูลภาษาอังกฤษที่รวบรวมไว้ได้ที่

http://www.arts.chula.ac.th/~ling/EngCorp/

อย่างไรก็ตาม แหล่งทรัพยากรที่รวบรวมไว้นั้นทำเมื่อปี 2541 ที่อยู่ที่ให้ไว้ในรายงานของบางเวบไซต์จึงอาจจะไม่ตรงตามจริงในปัจจุบัน หากเป็นเช่นนั้น ผู้ใช้ก็สามารถใช้โปรแกรมค้นหาโดยทั่วไป เช่น http://www.google.com/ เพื่อค้นหาตามชื่อของเวบไซต์ดังกล่าวได้

ท้ายที่สุด ผู้เขียนขอขอบคุณฝ่ายวิจัย คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยที่ได้ให้การสนับสนุนงานวิจัยชิ้นนี้ รวมทั้งสนับสนุนการตีพิมพ์หนังสือเล่มนี้ ขอขอบคุณ คุณดำเนิน การเด่น ที่ได้สละเวลาอันมีค่าอ่านและชี้ให้เห็นข้อผิดพลาดในรายงานฉบับแรก ผศ.ปรีมา มัลลิกะมาส ที่ได้อ่านและให้คำแนะนำเพื่อปรับปรุงแก้ไขในฉบับต่อมา อ.วิไลวรรณ อรุณมานะกุล ภรรยาของผู้วิจัยที่ได้ช่วยขัดเกลาสำนวน ให้ข้อคิดเห็นและกำลังใจตลอดมาจนหนังสือเล่มนี้เสร็จสมบูรณ์ขึ้น และขอขอบคุณ ผศ.ดร.เพียรศิริ วงศ์วิภานนท์ที่ได้สละเวลาอันมีค่าอ่าน ให้คำติชม และเขียนคำนิยมให้กับหนังสือเล่มนี้

หากมีข้อผิดพลาดประการใดในหนังสือเล่มนี้ ผู้เขียนขอน้อมรับข้อผิดพลาดและคำวิจารณ์หรือคำติชมใดใด เพื่อนำมาปรับปรุงหนังสือนี้ต่อไปใน

อนาคต แต่หากมีความดีใดใดจากหนังสือเล่มนี้ ผู้เขียนขอมอบแด่ รศ.ดร.ปราณี กุลละวณิชย์ นักภาษาศาสตร์คนแรกที่ผู้เขียนได้รู้จักและขอคำปรึกษาทางวิชาการ และ ผศ.ดร.เพียรศิริ วงศ์วิภานนท์ ซึ่งเป็นผู้ที่ชักจูงผู้เขียนเข้าสู่แวดวงวิชาการทางภาษาศาสตร์ ทั้งสองท่านเป็นอาจารย์ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยและเป็นแบบอย่างของอาจารย์อย่างที่ผู้เขียนอยากจะเป็นเสมอมา

วิโรจน์ อรุณมานะกุล
10 พ.ค. 2545

สารบัญ

บทที่ 1 คลังข้อมูลภาษา 1

1 คลังข้อมูลภาษาคืออะไร 1

2 ความเป็นมาของภาษาศาสตร์คลังข้อมูล 3

3 ขอบเขตของภาษาศาสตร์คลังข้อมูล 10

4 พัฒนาการของคลังข้อมูลภาษา 11

4.1 คลังข้อมูลยุคแรก 13

4.1.1 Brown Corpus 13

4.1.2 Lancaster-Oslo/Bergen (LOB) Corpus 14

4.1.3 London-Lund Corpus (LLC) 15

4.2 คลังข้อมูลยุคที่สอง 16

4.2.1 The Cobuild Project 16

4.2.2 The Longman Corpus Network 17

4.2.3 The British National Corpus 18

4.2.4 The International Corpus of English (ICE) 20

5. ประเภทของคลังข้อมูลภาษา 21

6. ความหมายของ CORPUS และ ARCHIVE 23

7. คำนวณภาพขนาดของคลังข้อมูล 24

8. ปัจจุบันและอนาคตของภาษาศาสตร์คลังข้อมูล 24

บทที่ 2 คลังข้อมูลกับการเรียนการสอนภาษา 27

1. บทนำ 27

2. การนำคลังข้อมูลภาษามาใช้เป็นหลักในการสอนภาษา 34

2.1 วิธีการเรียนการสอนด้วยการใช้คลังข้อมูลภาษา 34
2.2 บทบาทของผู้เรียนและผู้สอนในชั้นเรียน 36
3. การประยุกต์ใช้คลังข้อมูลภาษาในการสอนภาษา 37
บทที่ 3 คลังข้อมูลกับการวิจัยภาษา 45
1. บทนำ 45
2. การศึกษาระดับคำ 48
3. การศึกษาระดับโครงสร้าง 50
4. การศึกษาด้าน ESP และการแปรตามทำเนียบภาษา 51
5. หัวข้อในการวิจัยทางภาษาโดยใช้คลังข้อมูล 53
5.1 การศึกษาเรื่องคำ 53
5.2 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องคำและหน่วยคำ 54
5.3 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องประโยค 56
5.3.1 การศึกษาความยาวประโยค 56
5.3.2 การศึกษากระบวนการทางไวยากรณ์ 56
5.4 การศึกษาวัจนกรรมในภาษาพูด 58
5.5 การศึกษาด้านการแปรของภาษา 58
5.6 การศึกษาด้านวัจนลีลา 61
บทที่ 4 คลังข้อมูลกับการแปล 63
1. บทนำ 63
2. ความเป็นมาของศาสตร์การแปล 65
3. ทฤษฎีระบบหลายขั้ว 67
4. แนวคิดเรื่องรูปแบบปกติ 69
5. คลังข้อมูลภาษากับศาสตร์การแปลแบบพรรณา 70

6. ศาสตร์การแปลและการวิเคราะห์เปรียบต่าง 71
7. การศึกษาลักษณะเฉพาะของภาษาแปล 73
7.1 การศึกษาการเบี่ยงเบนจากระบบและรูปแบบปกติ 73
7.2 การศึกษาสากลลักษณ์ของการแปล 74
8. การใช้คลังข้อมูลภาษาเพื่อช่วยในการแปล 76
8.1 การใช้คลังข้อมูลในการสอนแปล 77
8.2 การใช้คลังข้อมูลเป็นทรัพยากรการแปล 83
8.2.1 การทดลองที่มหาวิทยาลัย Dublin City 91
8.2.2 การทดลองที่มหาวิทยาลัย Innsbruck 98
บทที่ 5 คลังข้อมูลกับการทำพจนานุกรม 101
1. บทนำ 101
2. พจนวิทยาและการทำพจนานุกรม 102
3. การทำพจนานุกรมแบบใช้คลังข้อมูลภาษา 105
4. อรรถศาสตร์คำศัพท์ 108
5, เครือข่ายคำ 110
6. ศัพทวิทยาและการทำประมวลศัพท์ 119
7. บทสรุป 122
บทที่ 6 คลังข้อมูลกับการประมวลผลภาษา 123
1. บทนำ 123
2. ความสัมพันธ์ระหว่างภาษาศาสตร์และสถิติ 123
3. การใช้สถิติในการแก้ปัญหาความกำกวม 132
4. บทสรุป 143

บทที่ 7 คลังข้อมูลภาษาผู้เรียน 145
1. บทนำ 145
2. การใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียน 149
2.1 ประโยชน์ในการศึกษาข้อผิด 150
2.2 ประโยชน์ในการศึกษาเปรียบต่าง 151
2.3 ประโยชน์ในการศึกษาวิเคราะห์การใช้ภาษา 153
2.4 ประโยชน์ด้านการจัดการเรียนการสอน 153
3. การพัฒนาคลังข้อมูลภาษาผู้เรียน 156
4. โครงการพัฒนาคลังข้อมูลภาษาผู้เรียนที่ต่างๆ 158
5. ตัวอย่างงานวิจัยที่ใช้คลังข้อมูลภาษาผู้เรียน 164
6. การกำกับข้อผิด 168
7. บทสรุป 174
บทที่ 8 การออกแบบและการสร้างคลังข้อมูลภาษา 177
1. ข้อควรพิจารณา 177
1.1 วัตถุประสงค์ 177
1.2 ลักษณะของคลังข้อมูลเป็นแบบสถิตย์หรือแบบพลวัต 178
1.3 การเป็นตัวแทนภาษาและความสมดุล 179
1.4 ขนาดของคลังข้อมูลภาษา 180
1.5 การป้อนข้อมูลเข้า 183
2. ขั้นตอนในการจัดสร้างคลังข้อมูลภาษา 184
2.1 การออกแบบโครงสร้าง 184
2.2 วางแผนการเก็บข้อมูล 185
2.3 ขออนุญาตเจ้าของลิขสิทธิ์ 186

2.4 การป้อนข้อมูลเข้า 188
2.5 การกำกับข้อมูล 189
3 ตัวอย่างการจัดสร้างคลังข้อมูลภาษา 189
3.1 คลังข้อมูล CSAE 190
3.1.1 แรงจูงใจ 190
3.1.2 ประโยชน์ที่คาดไว้ 190
3.1.3 สิ่งที่คำนึงถึงในการสร้างคลังข้อมูล CSAE 191
3.1.4 ขั้นตอนของการสร้าง CSAE 193
3.2 คลังข้อมูล BNC 194
3.2.1 ลักษณะทั่วไป 194
3.2.2 การคัดเลือกข้อมูล 196
3.2.3 การสร้างคลังข้อมูล BNC 198
บทที่ 9 การนำข้อมูลภาษาเข้าคอมพิวเตอร์ 201
1. การเก็บข้อมูลในคอมพิวเตอร์ 201
2. รหัสอักขระ 203
2.1 รหัสอักขระภาษาต่างๆ 204
2.2 ตัวอักขระภาษาไทย 206
2.3 รหัสยูนิโค้ด 208
3. ไฟล์ข้อมูลแบบต่างๆ 211
4. การแปลงไฟล์ข้อมูล 215
5. การใช้โปรแกรมรู้จำตัวอักษรและสแกนเน่อร์ 219
6. การใช้โปรแกรม HTTrack 220
6.1 ตัวเลือกต่างๆในโปรแกรม HTTrack 225

6.2 ข้อแนะนำในการใช้โปรแกรม HTTrack 235
7. พระราชบัญญัติลิขสิทธิ์พ.ศ. ๒๕๓๗ 236
บทที่ 10 การกำกับข้อมูล 241
1. เหตุผลที่มีการกำกับข้อมูล 241
2. การกำกับข้อมูลคืออะไร 242
3. หน้าที่และบทบาทของ TEI 243
4. SGML และ XML 245
4.1 องค์ประกอบของเอกสาร SGML/XML 248
4.1.1 SGML Declaration 248
4.1.2 XML Declaration 250
4.1.3 Document instance 250
4.1.4 Document Type Definition (DTD) 251
4.2 การกำหนดส่วนเฉพาะ (Marked Section) 259
4.3 ตัวอย่างของแฟ้มข้อมูลที่กำกับด้วย SGML และ XML 260
5. แท็กพื้นฐานที่กำหนดโดย TEI 263
5.1 โครงสร้างของเอกสารตามข้อกำหนด TEI 264
5.2 การกำกับข้อมูลส่วนเนื้อความ 265
5.3 การกำกับข้อมูลในคลังข้อมูลภาษา 267
5.4 การกำกับคลังข้อมูลภาษาตามแนวทางของ TEI 270
5.4.1 การกำกับข้อมูลบริบท 270
5.4.2 การกำกับข้อมูลทางภาษาศาสตร์ 279
6. สรุป 281

บทที่ 11 โปรแกรมคอนคอร์แด๊นซ์ 283
1. บทนำ 283
2. ประเภทของโปรแกรมคอนคอร์แด๊นซ์ 285
3. การค้นข้อมูล 286
4. การแสดงผล 287
5. ข้อจำกัดของโปรแกรมคอนคอร์แด๊นซ์ 288
6. ประโยชน์ของโปรแกรมคอนคอร์แด๊นซ์ 289
6.1 ประโยชน์ด้านการวิจัยทางภาษาศาสตร์ 289
6.2 ประโยชน์ด้านการสอนภาษา 290
6.3 ประโยชน์ในงานประยุกต์อื่นๆ 290
7. ข้อควรระวัง 292
8 ตัวอย่างการใช้โปรแกรมคอนคอร์แด๊นซ์ 296
8.1 การโหลดคลังข้อมูลเข้าโปรแกรม 297
8.2 การหารายการความถี่ของคำ 298
8.3 การค้นคำหลักพร้อมบริบท 299
8.4 การหาคำปรากฏร่วม 301
8.5 การค้นหากลุ่มคำ 304
8.6 การค้นแบบซับซ้อน 305
8.7 การแสดงตำแหน่งการปรากฏของคำ 308
8.8 การหาคำสำคัญเมื่อเทียบกับคลังข้อมูลอ้างอิง 309
8.9 การใช้ AntConc กับข้อมูล XML 311
8.10 การใช้ AntConc กับข้อมูลภาษาไทย 312
9. การใช้โปรแกรมคอนคอร์แด๊นซ์ภาษาไทย 314

10. การใช้คลังข้อมูลภาษาไทยแห่งชาติ 317
บทที่ 12 คำปรากฏร่วมและหน่วยเสมือนวลี 319
1. ความเป็นมา 319
2. ประเภทของคำปรากฏร่วม 322
3. เกณฑ์การพิจารณาคำปรากฏร่วม 323
4. การระบุหาคำปรากฏร่วม 324
4.1 การใช้ความถี่ 326
4.2 การใช้ค่าเฉลี่ยและความแปรปรวน 326
4.3. การทดสอบสมมติฐาน 328
4.3.1 การทดสอบ t-test 329
4.3.2 การทดสอบ Pearson's chi-square 332
4.3.3 การทดสอบ Berry-Rogghe's z-score 334
4.3.4 การทดสอบค่า Likelihood ratio 338
4.4 การใช้ค่า Mutual Information 339
4.5 การทดสอบแบบอื่นๆ 341
4.6 การหาคำปรากฏร่วมมากกว่าสองคำ 343
4.7 การหาคำปรากฏร่วมระหว่างสองภาษา 344
5. สรุป 345
บทที่ 13 การใช้เว็บเป็นคลังข้อมูลภาษา 347
1. บทนำ 347
2. ข้อมูลภาษาในเว็บ 347
3. การรวบรวมข้อมูลเว็บสร้างเป็นคลังข้อมูลภาษา 352
4. การใช้เว็บเป็นคลังข้อมูลภาษา 356

5. WEB-BASED คอนคอร์แด๊นซ์ 359
5.1 โปรแกรม WebCorp 361
5.2 โปรแกรม KWiCFinder 364
6 บทสรุป 366
บรรณานุกรม 367

คำนิยม

หนังสือ **ภาษาศาสตร์คลังข้อมูล** เล่มนี้ เป็นหนังสือที่มีคุณค่าและเหมาะสมแก่กาลเวลาเป็นอย่างยิ่ง ด้วยเหตุผลที่ว่าทั้งการศึกษาภาษาและการศึกษาด้านทฤษฎีภาษาศาสตร์ได้พัฒนามาถึงวาระที่เราไม่อาจจะอ้างอิงความถูกต้องของข้อสรุปหรือทฤษฎีที่นำเสนอ จากความรู้สึกหรือความสำนึกรู้ในฐานะเจ้าของภาษาเพียงอย่างเดียวได้อีกต่อไปแล้ว แต่จะต้องพยายามใช้ความเป็นจริงและข้อเท็จจริงของภาษามาเป็นหลักฐานสนับสนุนด้วย เครื่องมือหนึ่งที่จะช่วยให้เราเข้าถึงความเป็นจริงของภาษาได้ใกล้เคียงที่สุดก็คือคลังข้อมูลภาษา (Language corpus) ซึ่งรวบรวมข้อมูลภาษาที่ได้มีผู้ใช้ทั้งในการพูดและการเขียนมาเก็บรวบรวมไว้อย่างเป็นระบบเพื่อความสะดวกในการค้นหาข้อเท็จจริงที่จะนำไปพิสูจน์สมมุติฐาน ข้อสรุป หรือทฤษฎีของเรา ไม่ว่าเราจะทำงานด้านการศึกษาภาษามนุษย์ในภาพรวม ภาษาแต่ละภาษา ปรากฏการณ์ภาษา การแปลทั้งด้วยมือมนุษย์และด้วยเครื่อง การทำเครื่องมือช่วยในการสื่อภาษา เช่น พจนานุกรม และ อรรถาภิธาน ตลอดจนถึงการสร้างความสามารถทางด้านโปรแกรมที่ทำให้คอมพิวเตอร์สามารถประมวลผลภาษาธรรมชาติของมนุษย์ได้

วิวัฒนาการการสร้างและการใช้คลังข้อมูลภาษาดังกล่าวนี้ คงจะเกิดขึ้นไม่ได้หากไม่มีการนำวิวัฒนาการทางคอมพิวเตอร์มาผนวกเข้ากับวิวัฒนาการทางด้านภาษาศาสตร์ จนแตกเป็นแขนงขึ้นในสาขาภาษาศาสตร์ ที่เรียกขานกันว่า **ภาษาศาสตร์คอมพิวเตอร์** (Computational linguistics) และในสาขาวิทยาการคอมพิวเตอร์ ที่เรียกกันโดยรวมว่า **การประมวลภาษาธรรมชาติ** (Natural language processing) การผนวกกันของทั้งสองสาขาวิชาการนี้เป็นเรื่องที่ค่อนข้างยากลำบากมาแต่ต้น เนื่องจากมีนักวิชาการไม่มากคนนักที่จะ

เชี่ยวชาญทั้งด้านภาษาศาสตร์และด้านคอมพิวเตอร์ อาจารย์ ดร.วิโรจน์ อรุณมานะกุล เป็นหนึ่งในผู้ที่มีความรู้อย่างลึกซึ้งในทั้งสองสาขาวิชาการ เนื่องด้วยเริ่มต้นมาจากพื้นฐานของนักคอมพิวเตอร์ในระดับปริญญาบัณฑิต และมาแตกแขนงเป็นวิชาภาษาศาสตร์ในระดับปริญญามหาบัณฑิต และต่อยอดอย่างสมบูรณ์ด้วยปริญญาดุษฎีบัณฑิตทางด้านภาษาศาสตร์คอมพิวเตอร์

ความเป็นนักวิชาการที่เชี่ยวชาญทั้งสาขาภาษาศาสตร์และสาขาคอมพิวเตอร์เพียงอย่างเดียว คงไม่เพียงพอที่จะทำให้หนังสือ **ภาษาศาสตร์คลังข้อมูล** นี้เกิดประโยชน์แก่นิสิตนักศึกษาและผู้สนใจได้อย่างเต็มที่ หากขาดเสียซึ่งความเป็นครูของผู้เขียน อาจารย์ ดร.วิโรจน์ อรุณมานะกุล เป็นครูที่ดีมาโดยตลอด นอกจากภูมิความรู้จะแน่นแล้ว ยังสอนหนังสือจากมุมของลูกศิษย์ผู้ยังไม่รู้มากนัก มิใช่จากมุมของความเป็นผู้เชี่ยวชาญของตน หนังสือเล่มนี้เขียนอย่างครูเขียนจึงอ่านง่ายและไม่ทำให้คนหน้าใหม่ที่อยากรู้หรือเพิ่งเริ่มศึกษาเรื่องนี้เกิดความรู้สึกระย่อเสียก่อน ส่วนผู้ที่มีความรู้อยู่บ้างแล้วไม่ว่าจะเป็นนักภาษา นักภาษาศาสตร์ หรือนักคอมพิวเตอร์ ก็สามารถจะได้ประโยชน์จากความสมบูรณ์ของเนื้อหาและแหล่งอ้างอิง ที่จะนำไปใช้ในการพัฒนาด้วยตนเองต่อไป

ขอแสดงความยินดีและความขอบคุณอาจารย์ ดร.วิโรจน์ อรุณมานะกุล ที่ได้เขียนหนังสือเล่มนี้ขึ้นมา เพื่อให้เป็นประโยชน์แก่ทุกคนที่สนใจเรื่องการศึกษาภาษาด้วยข้อมูลและหลักฐานอ้างอิงที่เป็นข้อเท็จจริง ไม่ว่าจะมาจากภูมิหลังอย่างใดและมีวัตถุประสงค์อย่างใดในการศึกษาภาษาของตน

เพียรศิริ วงศ์วิภานนท์ (Ph.D.)

บทนำ

หนังสือภาษาศาสตร์คลังข้อมูลนี้ เขียนขึ้นเพื่อให้ความรู้พื้นฐานต่างๆ ที่เกี่ยวข้องกับคลังข้อมูลภาษา เพื่อให้ผู้อ่านได้เข้าใจและเห็นความสำคัญของการนำคลังข้อมูลภาษาไปใช้ประโยชน์ในการศึกษาด้านต่างๆ หนังสือเล่มนี้มีจุดเริ่มต้นมาจากรายงานการวิจัยเรื่อง “แนวทางการพัฒนาคลังข้อมูลภาษา: กรณีศึกษาจากการสร้างคลังข้อมูลภาษาอังกฤษ” ซึ่งผู้เขียนได้รับทุนสนันสนุนการวิจัยจากฝ่ายวิจัย คณะอักษรศาสตร์ ในปีพ.ศ. 2541 โครงการวิจัยนี้เป็นโครงการที่ศึกษากลวิธีที่ใช้ในการสร้างคลังข้อมูลภาษา โดยการสำรวจคลังข้อมูลภาษาจากแหล่งต่างๆ เพื่อวิเคราะห์ถึงลักษณะความแตกต่างรวมทั้งปัจจัยที่กำหนดลักษณะการจัดวางโครงสร้างคลังข้อมูลแต่ละรูปแบบ และสำรวจโปรแกรมซอฟท์แวร์ที่ใช้ร่วมกับคลังข้อมูลภาษาเหล่านั้น พร้อมทั้งทดลองสร้างคลังข้อมูลภาษา โดยคัดเลือกเอกสารอิเล็กทรอนิกส์ภาษาอังกฤษที่สามารถสืบค้นได้ทางอินเตอร์เน็ตและไม่มีข้อจำกัดเรื่องลิขสิทธิ์ เพื่อนำมารวบรวมเป็นคลังข้อมูลภาษาอังกฤษเพื่อใช้ประโยชน์ด้านการเรียนการสอนและงานวิจัยในคณะอักษรศาสตร์

หนังสือเล่มนี้ปรับปรุงขึ้นจากตอนที่หนึ่งของรายงานวิจัยดังกล่าว ซึ่งเป็นตอนที่เขียนขึ้นเพื่อให้พื้นฐานความรู้ที่จำเป็นทางด้านภาษาศาสตร์คลังข้อมูล ภาษาศาสตร์คลังข้อมูลเป็นวิชาที่ได้รับความสนใจอย่างมากในช่วงทศวรรษที่ 1990 นี้ โดยเน้นที่การศึกษาภาษาในแง่มุมต่างๆ โดยใช้ประโยชน์จากการรวบรวมข้อมูลภาษาจำนวนมาก คลังข้อมูลภาษาที่รวบรวมได้นี้สามารถใช้ประโยชน์ได้ทั้งในแง่ของการเรียนการสอนภาษา การวิจัยภาษา การแปล การจัดทำพจนานุกรม และการประมวลผลภาษาธรรมชาติ

บทที่ 1 เป็นบทแนะนำให้รู้ว่าคลังข้อมูลภาษาคืออะไร เพราะเหตุใดจึงได้รับความสนใจจากนักภาษาศาสตร์ และนักคอมพิวเตอร์เป็นอย่างมาก ซึ่งจะขยายความต่อในบทที่ 2 โดยจะกล่าวถึงประโยชน์ของคลังข้อมูลที่มีต่องานด้านต่างๆ สี่ด้าน ได้แก่ งานด้านการเรียนการสอนภาษา งานด้านการศึกษาวิจัยทางภาษาศาสตร์ งานด้านการแปล และงานทางด้านการประมวลผลภาษาธรรมชาติ เมื่อผู้อ่านเห็นภาพว่าคลังข้อมูลภาษาคืออะไรและมีประโยชน์อย่างไรแล้ว จึงจะได้กล่าวถึงการออกแบบและจัดสร้างคลังข้อมูลภาษาในบทที่ 3 โดยจะกล่าวถึงปัจจัยต่างๆที่ควรคำนึงในการออกแบบคลังข้อมูล ส่วนในบทที่ 4 เป็นเรื่องของการกำกับข้อมูลและมาตรฐานของการกำกับข้อมูล โดยจะกล่าวถึง SGML เป็นหลักเนื่องจากเป็นมาตรฐานที่ได้รับการยอมรับในปัจจุบันให้เป็นมาตรฐานของการกำกับข้อมูลในคลังข้อมูลภาษา แม้ว่าในปัจจุบัน จะมีผู้กล่าวถึงการใช้ XML ในการกำกับข้อมูลต่างๆ แต่เนื่องจาก SGML เป็นต้นกำเนิดของมาตรฐาน และ XML ก็ถือเป็น SGML ที่ถูกทำให้ง่ายขึ้น ผู้ที่เข้าใจ SGML แล้วก็จะสามารถศึกษา XML ได้โดยง่าย ในที่นี้ จึงได้นำเสนอเรื่องของ SGML เป็นสำคัญ

เนื่องจากการที่จะใช้ประโยชน์จากคลังข้อมูลภาษาให้ได้เต็มที่นั้น จำเป็นต้องอาศัยโปรแกรมหรือซอฟท์แวร์เพื่อช่วยในการค้น วิเคราะห์ หรือจัดสร้างคลังข้อมูล ในบทที่ 5 จึงกล่าวถึงโปรแกรมคอนคอร์แดนซ์ซึ่งเป็นโปรแกรมพื้นฐานสำหรับใช้ร่วมกับคลังข้อมูลภาษา เพื่อช่วยในการค้นหาคำและบริบทที่ปรากฏร่วมกับคำนั้นๆ โดยจะกล่าวถึงลักษณะทั่วไปของโปรแกรมประเภทนี้ และนำเสนอลักษณะเฉพาะของโปรแกรมคอนคอร์แดนซ์จำนวนหนึ่ง ส่วนบทที่ 6 จะกล่าวถึงวิธีการทางสถิติที่สามารถนำมาประยุกต์ใช้เพื่อช่วยหาคำที่มักปรากฏใช้ร่วมกัน ซึ่งเป็นการใช้ประโยชน์จากคลังข้อมูลภาษาในอีกลักษณะหนึ่ง

สำหรับส่วนที่สองและส่วนที่สามของงานวิจัยดังกล่าว ผู้เขียนไม่ได้นำมาเสนอในหนังสือเล่มนี้ เนื่องจากเป็นการรวบรวมแหล่งทรัพยากรที่เกี่ยวข้องกับคลังข้อมูลภาษาอังกฤษ และการจัดสร้าง CU English Text Archive ตามลำดับ ซึ่งไม่ได้เป็นส่วนที่ให้ความรู้พื้นฐานเกี่ยวกับศาสตร์ด้านนี้โดยตรง ผู้สนใจสามารถดาว์นโหลดรายงานวิจัยในส่วนนี้ได้จากอินเตอร์เน็ตที่

http://www.arts.chula.ac.th/~ling/paper/corp2.zip

http://www.arts.chula.ac.th/~ling/paper/corp3.zip

และสามารถใช้คลังข้อมูลภาษาอังกฤษที่รวบรวมไว้ได้ที่

http://www.arts.chula.ac.th/~ling/EngCorp/

อย่างไรก็ตาม แหล่งทรัพยากรที่รวบรวมไว้นั้นทำเมื่อปี 2541 ที่อยู่ที่ให้ไว้ในรายงานของบางเวบไซต์จึงอาจจะไม่ตรงตามจริงในปัจจุบัน หากเป็นเช่นนั้น ผู้ใช้ก็สามารถใช้โปรแกรมค้นหาโดยทั่วไป เช่น http://www.google.com/ เพื่อค้นหาตามชื่อของเวบไซต์ดังกล่าวได้

ท้ายที่สุด ผู้เขียนขอขอบคุณฝ่ายวิจัย คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยที่ได้ให้การสนับสนุนงานวิจัยชิ้นนี้ รวมทั้งสนับสนุนการตีพิมพ์หนังสือเล่มนี้ ขอขอบคุณ คุณดำเนิน การเด่น ที่ได้สละเวลาอันมีค่าอ่านและชี้ให้เห็นข้อผิดพลาดในรายงานฉบับแรก ผศ.ปรีมา มัลลิกะมาส ที่ได้อ่านและให้คำแนะนำเพื่อปรับปรุงแก้ไขในฉบับต่อมา อ.วิไลวรรณ อรุณมานะกุล ภรรยาของผู้วิจัยที่ได้ช่วยขัดเกลาสำนวน ให้ข้อคิดเห็นและกำลังใจตลอดมาจนหนังสือเล่มนี้เสร็จสมบูรณ์ขึ้น และขอขอบคุณ ผศ.ดร.เพียรศิริ วงศ์วิภานนท์ที่ได้สละเวลาอันมีค่าอ่าน ให้คำติชม และเขียนคำนิยมให้กับหนังสือเล่มนี้

หากมีข้อผิดพลาดประการใดในหนังสือเล่มนี้ ผู้เขียนขอน้อมรับข้อผิดพลาดและคำวิจารณ์หรือคำติชมใดใด เพื่อนำมาปรับปรุงหนังสือนี้ต่อไปใน

อนาคต แต่หากมีความดีใดใดจากหนังสือเล่มนี้ ผู้เขียนขอมอบแด่ รศ.ดร.ปราณี กุลละวณิชย์ นักภาษาศาสตร์คนแรกที่ผู้เขียนได้รู้จักและขอคำปรึกษาทางวิชาการ และ ผศ.ดร.เพียรศิริ วงศ์วิภานนท์ ซึ่งเป็นผู้ที่ชักจูงผู้เขียนเข้าสู่แวดวงวิชาการทางภาษาศาสตร์ ทั้งสองท่านเป็นอาจารย์ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยและเป็นแบบอย่างของอาจารย์อย่างที่ผู้เขียนอยากจะเป็นเสมอมา

วิโรจน์ อรุณมานะกุล
10 พ.ค. 2545

สารบัญ

บทที่ 1 คลังข้อมูลภาษา **1**
1 คลังข้อมูลภาษาคืออะไร 1
2 ความเป็นมาของภาษาศาสตร์คลังข้อมูล 3
3 ขอบเขตของภาษาศาสตร์คลังข้อมูล 8
4 พัฒนาการของคลังข้อมูลภาษา 9
4.1 คลังข้อมูลยุคแรก 11
4.1.1 Brown Corpus 11
4.1.2 Lancaster-Oslo/Bergen (LOB) Corpus 12
4.1.3 London-Lund Corpus (LLC) 13
4.2 คลังข้อมูลยุคที่สอง 13
4.2.1 The Cobuild Project 14
4.2.2 The Longman Corpus Network 15
4.2.3 The British National Corpus 16
4.2.4 The International Corpus of English (ICE) 18
5. ประเภทของคลังข้อมูลภาษา 18
6. ความหมายของ CORPUS และ ARCHIVE 20
7. คำนวณภาพขนาดของคลังข้อมูล 21
8. ปัจจุบันและอนาคตของภาษาศาสตร์คลังข้อมูล 22

บทที่ 2 ประโยชน์ของคลังข้อมูลภาษา **24**
1. บทนำ 24
2. ประโยชน์ด้านการเรียนการสอนภาษา 26

2.1 การนำคลังข้อมูลภาษามาใช้เป็นหลักในการสอนภาษา 33
2.1.1 วิธีการเรียนการสอนด้วยการใช้คลังข้อมูลภาษา 33
2.1.2 บทบาทของผู้เรียนและผู้สอนในชั้นเรียน 34
2.2. การประยุกต์ใช้คลังข้อมูลภาษาในการสอนภาษา 36
3. ประโยชน์ด้านการวิจัยทางภาษาศาสตร์ 44
3.1 การศึกษาระดับคำ 47
3.2 การศึกษาระดับโครงสร้าง 49
3.3 การศึกษาด้าน ESP และการแปรตามทำเนียบภาษา 50
4. หัวข้อในการวิจัยทางภาษาศาสตร์ด้วยคลังข้อมูลภาษา 52
4.1 การศึกษาเรื่องคำ 52
4.2 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องคำและหน่วยคำ 52
4.3 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องประโยค 55
4.3.1 การศึกษาความยาวประโยค 55
4.3.2 การศึกษากระบวนการทางไวยากรณ์ 55
4.4 การศึกษาวัจนกรรมในภาษาพูด 56
4.5 การศึกษาด้านการแปรของภาษา 57
5. ประโยชน์ด้านการแปลภาษา 59
5.1 ความเป็นมาของศาสตร์การแปล 62
5.2 ทฤษฎีระบบหลายขั้ว 64
5.3 แนวคิดเรื่องรูปแบบปกติ 66
5.4 คลังข้อมูลภาษากับศาสตร์การแปลแบบพรรณา 66
5.5 ศาสตร์การแปลและการวิเคราะห์เปรียบต่าง 68
5.6 การศึกษาลักษณะเฉพาะของภาษาแปล 69

5.6.1 การศึกษาการเบี่ยงเบนจากระบบและรูปแบบปกติ 70
5.6.2 การศึกษาสากลลักษณ์ของการแปล 70
5.7 การใช้คลังข้อมูลภาษาเพื่อช่วยในการแปล..... 73
5.7.1 การใช้คลังข้อมูลในการสอนแปล..... 74
5.7.2 การใช้คลังข้อมูลเป็นทรัพยากรการแปล..... 80
6 ประโยชน์ด้านการประมวลผลภาษาธรรมชาติ..... 96
6.1 ความสัมพันธ์ระหว่างภาษาศาสตร์และสถิติ..... 97
6.2 การใช้สถิติในการแก้ปัญหาความกำกวม 106
7. สรุป 116

บทที่ 3 การออกแบบและจัดสร้างคลังข้อมูลภาษา 117
1. ข้อควรพิจารณา 117
1.1 วัตถุประสงค์..... 117
1.2 ลักษณะของคลังข้อมูลเป็นแบบสถิตย์หรือแบบพลวัต..... 118
1.3 การเป็นตัวแทนภาษาและความสมดุล 119
1.4 ขนาดของคลังข้อมูลภาษา..... 120
1.5 การป้อนข้อมูลเข้า..... 123
2. ขั้นตอนในการจัดสร้างคลังข้อมูลภาษา..... 124
2.1 การออกแบบโครงสร้าง..... 124
2.2 วางแผนการเก็บข้อมูล..... 125
2.3 ขออนุญาตเจ้าของลิขสิทธิ์ 126
2.4 การป้อนข้อมูลเข้า..... 128
2.5 การกำกับข้อมูล 129

3 ตัวอย่างการจัดสร้างคลังข้อมูลภาษา....129
3.1 คลังข้อมูล CSAE....130
3.1.1 แรงจูงใจ....130
3.1.2 ประโยชน์ที่คาดไว้....130
3.1.3 สิ่งที่คำนึงถึงในการสร้างคลังข้อมูล CSAE....131
3.1.4 ขั้นตอนของการสร้าง CSAE....133
3.2 คลังข้อมูล BNC....134
3.2.1 ลักษณะทั่วไป....134
3.2.2 การคัดเลือกข้อมูล....135
3.2.3 การสร้างคลังข้อมูล BNC....138

บทที่ 4 การกำกับข้อมูลในคลังข้อมูลภาษา....141
1. เหตุผลที่มีการกำกับข้อมูล....141
2. การกำกับข้อมูลคืออะไร....144
3. หน้าที่และบทบาทของ TEI....145
4. SGML....146
4.1 องค์ประกอบของเอกสาร SGML....148
4.1.1 SGML Declaration....148
4.1.2 Document instance....150
4.1.2 Document Type Definition (DTD)....151
4.2 การกำหนดโครงสร้างซ้อน....161
4.3 การกำหนดส่วนเฉพาะ....163
4.4 ตัวอย่างของแฟ้มข้อมูลที่กำกับด้วย SGML....164

5. แท็กพื้นฐานที่กำหนดโดย TEI 166
5.1 โครงสร้างของเอกสารตามข้อกำหนด TEI 167
5.2 การกำกับข้อมูลส่วนเนื้อความ..... 169
5.3 การกำกับข้อมูลในคลังข้อมูลภาษา 171
5.4 การกำกับคลังข้อมูลภาษาตามแนวทางของ TEI..... 173
5.4.1 การกำกับข้อมูลบริบท 174
5.4.2 การกำกับข้อมูลทางภาษาศาสตร์..... 182
6. สรุป 184

บทที่ 5 โปรแกรมคอนคอร์แดนซ์ 186
1. บทนำ 186
2. ประเภทของโปรแกรมคอนคอร์แดนซ์ 188
3. การค้นข้อมูล..... 189
4. การแสดงผล 190
5. ข้อจำกัดของโปรแกรมคอนคอร์แดนซ์ 191
6. ประโยชน์ของโปรแกรมคอนคอร์แดนซ์..... 192
6.1 ประโยชน์ด้านการวิจัยทางภาษาศาสตร์..... 192
6.2 ประโยชน์ด้านการสอนภาษา..... 193
6.3 ประโยชน์ในงานประยุกต์อื่นๆ 193
7. ข้อควรระวัง 195
8 ตัวอย่างโปรแกรมคอนคอร์แดนซ์..... 199
8.1 โปรแกรม Concordancer for Windows..... 200
8.2 โปรแกรม Word Smith 205

8.3 โปรแกรม MonoConc 208
8.4 โปรแกรม TACT 211

บทที่ 6 คำปรากฏร่วม 213
1. บทนำ 213
2. การใช้ความถี่ 217
3 การใช้ค่าเฉลี่ยและความแปรปรวน 218
4. การทดสอบสมมติฐาน 219
4.1 การทดสอบ t-test 220
4.2 การทดสอบ Pearson's chi-square 223
4.3 การทดสอบ Berry-Rogghe's z-score 225
4.4 การทดสอบค่า Likelihood ratio 229
5 การใช้ค่า mutual information 230
6. การทดสอบแบบอื่นๆ 232
7. คำปรากฏร่วมของคำมากกว่าสองคำ 234
8. คำปรากฏร่วมระหว่างสองภาษา 235
9. สรุป 236

ภาคผนวก
ทรัพยากรคลังข้อมูลภาษาบนอินเตอร์เน็ต 237
บรรณานุกรม 239

บทที่ 1 คลังข้อมูลภาษา

1. คลังข้อมูลภาษาคืออะไร

กล่าวโดยคร่าว คลังข้อมูลภาษา (corpus) คือ ข้อมูลภาษาเขียนหรือภาษาพูดที่เป็นภาษาที่ใช้จริงซึ่งถูกรวบรวมขึ้นมาในปริมาณที่มากเพียงพอตามข้อกำหนดหรือเงื่อนไขที่กำหนดขึ้น เพื่อนำคลังข้อมูลนั้นมาใช้ประโยชน์ในการศึกษาเรื่องต่างๆที่เกี่ยวข้องกับภาษา Crysta (1991) ได้ให้คำนิยามรวมทั้งอธิบายลักษณะของคลังข้อมูลภาษา ไว้ดังนี้

> **corpus (corpora)** A collection of LINGUISTIC DATA, either written texts or a TRANSCRIPTION of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a LANGUAGE (corpus linguistics). Linguistic DESCRIPTIONS which are 'corpus-restricted' have been the subject of criticism, especially by GENERATIVE GRAMMARIANS, who point to the limitation of corpora (e.g. that they are samples of PERFORMANCE only, and that only one still needs a means of PROJECTING beyond the corpus to be the language as a whole). In fieldwork on a new language, or in HISTORICAL study, it may be very difficult to get beyond one's corpus (i.e. it is a 'closed' as opposed to an 'extendable' corpus), but in languages where linguists have regular access to NATIVE-SPEAKERS (and may be native-speakers

themselves) their approach will invariably be 'corpus-based', rather than corpus-restricted. Corpora provide the basis for one kind of COMPUTATIONAL LINGUISTICS. See Lyons 1968: Ch. 4.

(Crystal 1991. A Dictionary of Linguistics and Phonetics.)

ดังนั้น คลังข้อมูลภาษาจึงหมายถึงข้อมูลภาษาเขียนหรือภาษาพูดที่รวบรวมขึ้นมาเป็นจำนวนที่มากเพียงพอสำหรับใช้ในการศึกษาด้านภาษา ในปัจจุบัน เมื่อกล่าวถึงคลังข้อมูลภาษาจะหมายถึงข้อมูลภาษาที่ได้มีการเก็บบันทึกไว้ในระบบคอมพิวเตอร์ เนื่องจากคลังข้อมูลภาษาเป็นแหล่งที่ให้ข้อมูลการใช้ภาษาที่เกิดขึ้นจริง คลังข้อมูลภาษาจึงเป็นแหล่งข้อมูลที่สำคัญต่อการศึกษาวิจัยทางด้านภาษาและการสอนภาษานั้นๆ แนวคิดในการใช้ข้อมูลภาษาที่เกิดขึ้นจริงเพื่อการศึกษาทางภาษาศาสตร์ ได้เริ่มมีมาตั้งแต่ในช่วงทศวรรษที่ 50 อย่างไรก็ตามการเก็บข้อมูลภาษาจำนวนมากและการใช้ประโยชน์จากข้อมูลภาษาในยุคนั้นก็ยังเป็นงานที่ยากลำบากและใช้เวลามากเนื่องจากจำเป็นต้องอาศัยแรงงานคนเป็นหลัก ต่อมาเมื่อมีการใช้คอมพิวเตอร์แพร่หลายมากขึ้น การสร้างและใช้ประโยชน์จากคลังข้อมูลจึงเป็นไปได้โดยสะดวกมากขึ้น และเพราะการพัฒนาประสิทธิภาพของระบบคอมพิวเตอร์ทั้งในแง่ความเร็วในการทำงานและความสามารถในการเก็บข้อมูลขนาดใหญ่นี้เอง เราจึงสามารถใช้คอมพิวเตอร์ในการค้นหาคำหรือข้อความในคลังข้อมูลขนาดใหญ่ได้อย่างแม่นยำและรวดเร็ว สามารถใช้คอมพิวเตอร์เพื่อนับความถี่ คำนวณค่าทางสถิติ และจัดเรียงข้อมูลในรูปแบบที่ต้องการได้ คลังข้อมูลภาษาในปัจจุบันจะมีขนาดใหญ่มาก เช่น คลังข้อมูล BNC (British National Corpus) ซึ่งมีขนาดหนึ่งร้อยล้านคำ เมื่อ

เปรียบเทียบกับขนาดของคลังข้อมูลภาษาในยุคแรกๆ เช่น คลังข้อมูล Brown (Brown corpus) ซึ่งมีขนาดเพียงหนึ่งล้านคำ

2. ความเป็นมาของภาษาศาสตร์คลังข้อมูล

แม้ว่าความหมายของคลังข้อมูลภาษาในปัจจุบันจะหมายถึงข้อมูลที่จัดเก็บในคอมพิวเตอร์ แต่ความจริงมีการใช้และพูดถึงคลังข้อมูลภาษามาก่อนที่จะมีคอมพิวเตอร์ใช้กัน W. Nelson Francis (1992) พูดถึง Language corpora B.C. โดยที่ B.C. ย่อมาจาก before computer ว่ามีงานจำนวนมากที่ทำโดยอาศัยการรวบรวมข้อมูลภาษามาเพื่อการวิเคราะห์ เช่น การจัดทำพจนานุกรมภาษาอังกฤษ Oxford English Dictionary ซึ่งออกฉบับแรกในปีค.ศ.1928 และการเขียนหนังสือไวยากรณ์ Modern English Grammar ของ Jespersen ซึ่งเป็นงานที่ใช้ข้อมูลภาษาที่จดบันทึกมาไม่ได้เป็นข้อมูลที่จัดเก็บลงในคอมพิวเตอร์

ส่วนคำว่า corpus linguistics หรือภาษาศาสตร์คลังข้อมูล ก็เป็นคำใหม่ที่มีการใช้เมื่อไม่นานนี้ในช่วงทศวรรษ 1980 โดยพบคำนี้ในชื่อหนังสือต่างๆ เช่น Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research (edited by Jan Aarts and Willem Meijs, 1984), Corpus Linguistics Hard and Soft (edited by Merja Kytö et al. 1988), English Corpus Linguistics (edited by Karin Aijmer and Bengt Altenberg, 1991) เป็นต้น คำว่าภาษาศาสตร์คลังข้อมูลไม่ได้มีความหมายในลักษณะเดียวกับ ภาษาศาสตร์สังคม ภาษาศาสตร์จิตวิทยา เพราะไม่ได้จัดเป็นแขนงย่อยของภาษาศาสตร์ที่ศึกษาเรื่องเฉพาะในสาขา แต่เป็นเรื่องของการใช้ข้อมูลภาษาเป็นหลักสำคัญในการศึกษาภาษาในแง่มุมใดก็ได้ ซึ่งจะครอบคลุมการออกแบบ การสร้าง และการใช้คลังข้อมูลภาษาเพื่อตอบคำถามต่างๆ เกี่ยวกับภาษา

นักภาษาศาสตร์ที่ศึกษาภาษาโดยอิงกับข้อมูลภาษาจริงหรือที่เรียกว่า corpus linguists มักถูกนำมาเปรียบเทียบกับนักภาษาศาสตร์ที่ทำงานโดยอาศัยข้อมูลที่นึกคิดเอาหรือที่เรียกว่านักภาษาศาสตร์นั่งโต๊ะ (armchair linguists) โดยมีการพูดล้อเลียนว่า นักภาษาศาสตร์นั่งโต๊ะจะใช้เวลานั่งหลับตาครุ่นคิดอยู่บนเก้าอี้ของตนสักพักใหญ่ ๆ แล้วก็อุทานขึ้นว่า "*ใช่แล้วช่างเป็นข้อมูลที่วิเศษจริง*" แล้วก็จดบันทึกข้อมูลประโยคที่คิดขึ้นได้ จากนั้นก็นั่งหลับตาครุ่นคิดเช่นนั้นต่อไปอีก ในขณะที่นักภาษาศาสตร์คลังข้อมูลก็ถูกล้อเลียนว่า วันๆ เอาแต่นั่งอยู่กับกองข้อมูลมหาศาลของตน นั่งจัดลำดับข้อมูลที่มี นับความถี่ของคำ หาว่าหมวดคำของคำแรกในประโยคเป็นอะไรได้บ้าง เป็นต้น (Fillmore 1992 อ้างใน Svartvik 2007)

การศึกษาภาษาโดยอาศัยข้อมูลภาษาจริงอาจมองได้ว่าเริ่มต้นเป็นกระแสหลักของภาษาศาสตร์ในยุคของการศึกษาภาษาศาสตร์ของนักไวยากรณ์โครงสร้าง (structuralists) หลังทศวรรษที่ 1920 ซึ่งนักไวยากรณ์โครงสร้างจะให้ความสำคัญกับข้อมูลภาษาและถือว่าเป็นสิ่งจำเป็นสำหรับการศึกษาภาษา ทั้งนี้เนื่องจากนักไวยากรณ์โครงสร้างมองว่า การศึกษาทางภาษาควรเป็นการศึกษาในเชิงประจักษ์ (empiricism) ด้วยการสังเกตุจากข้อมูลภาษาจริงและอธิบายปรากฏการณ์ทางภาษาที่พบในลักษณะที่เป็นการศึกษาเชิงวิทยาศาสตร์ โดยจะไม่ยอมรับผลของการศึกษาภาษาศาสตร์ที่ผู้ศึกษานึกคิดคาดคะเนเอง อย่างไรก็ตาม แนวโน้มการศึกษาทางภาษาศาสตร์ในยุคต่อมาได้เบี่ยงเบนความสำคัญจากการใช้ข้อมูลภาษาจริงมาเป็นการใช้ความหยั่งรู้ (intuition) ของเจ้าของภาษาเป็นหลัก โดยบุคคลที่เสนอแนวคิดนี้คือ Noam Chomsky ในช่วงทศวรรษที่ 1950 ซึ่งชี้ว่ามีความแตกต่างระหว่างกฤตกรรมภาษา (performance) ซึ่งเป็นเรื่องของภาษาที่เกิดขึ้นที่เราได้เห็นได้ยิน กับสามัตถิยะภาษา (competence) ซึ่ง

เป็นเรื่องของความรู้ทางภาษาที่ทำให้เราสามารถตัดสินได้ว่าประโยคใดถูกไวยากรณีหรือไม่ นับจากนั้นมา ทิศทางของภาษาศาสตร์ก็เปลี่ยนจากการศึกษาแบบเชิงประจักษ์ไปสู่การศึกษาแบบเชิงเหตุผล (rationalism)

Chomsky เสนอแนวคิดว่าการศึกษาทางภาษาศาสตร์ไม่ควรยึดติดอยู่กับการใช้ข้อมูลภาษาจริง เพราะผลการศึกษาที่ได้จะเป็นเพียงการจัดระบบระเบียบข้อมูลที่รวบรวมมาได้ ซึ่งไม่ได้สะท้อนให้เห็นกลไกทางภาษาที่มนุษย์มีอยู่ภายใน Chomsky เสนอว่านักภาษาศาสตร์ควรมุ่งศึกษาสิ่งที่เป็นตัวความรู้ความสามารถทางภาษา (knowledge of language) ของมนุษย์หรือสามัตถิยะภาษามากกว่า คลังข้อมูลภาษาเป็นเพียงข้อมูลการปรากฏของรูปภาษาที่เจ้าของภาษานั้นใช้ และเป็นเพียงวัตถุดิบที่เราจะนำมาใช้เพื่ออนุมานหาสิ่งที่เป็นแบบจำลองของภาษาต่อไป Chomsky ชี้ประเด็นว่ามีความแตกต่างของระดับความรู้หรือทฤษฎีภาษาที่ได้จากวิธีการศึกษาภาษาที่ต่างกัน การจัดกลุ่มจัดระเบียบข้อมูลและหาข้อสรุปจากข้อมูลถูกมองว่าเป็นการได้มาซึ่งทฤษฎีที่ถูกต้องในระดับสังเกตุเห็น (observation adequacy) กับระดับพรรณา (descriptive adequacy) เท่านั้น เพราะเป็นการบอกได้ว่าประโยคแบบไหนที่พบในข้อมูลบ้างตามที่ค้นพบจริง และให้คำอธิบายความรู้เพิ่มเติมจากข้อมูลที่พบว่าทำไมประโยคไหนจึงถูกไวยากรณ์หรือไม่ถูกไวยากรณ์ แต่ก็ไม่ช่วยให้เราเข้าถึงทฤษฎีภาษาในระดับอธิบายได้ (explanatory adequacy) เช่นที่นักไวยากรณ์เพิ่มพูนในสาย Chomsky พยายามทำ คือ สร้างแบบจำลองที่จะอธิบายถึงความรู้ภาษาที่มนุษย์มีร่วมกันมาแก่กำเนิดหรือไวยากรณ์สากล (universal grammar) อันเป็นความรู้ภาษาที่ทำให้มนุษย์สามารถเรียนภาษาใดก็ได้เมื่อเติบโตในบริบทภาษานั้นๆ นอกจากนี้ ชอมสกี้ยังอ้างถึงเหตุผลต่างๆเพื่อชี้ให้เห็นว่าการใช้คลังข้อมูล

ภาษาไม่สามารถเป็นประโยชน์ต่อการสร้างแบบจำลองของสามัตถิยะภาษา เหตุผลต่างๆที่ยกมาสนับสนุนแนวคิดนี้ ได้แก่

1. จำนวนของประโยคในภาษานั้นมีมากมายมหาศาลและมีปริมาณที่ไม่จำกัด (non-finite) จึงเป็นไปไม่ได้ที่จะรวบรวมประโยคทั้งหมดมาศึกษา และการที่สมองมนุษย์มีขอบเขตที่จำกัด แสดงว่าสิ่งที่มีอยู่ในสมองมนุษย์คือกฎเกณฑ์หรือไวยากรณ์ทางภาษาเพราะเป็นสิ่งที่มีขอบเขตอันจำกัดได้ จึงควรสนใจศึกษาเรื่องไวยากรณ์มากกว่า

2. ความหยั่งรู้ (intuition) เท่านั้นที่จะสามารถใช้ตัดสินว่าประโยคใดไม่ถูกต้องตามไวยากรณ์ (ungrammatical)[1] การใช้เพียงคลังข้อมูลภาษาอย่างเดียวไม่สามารถทำให้เราเห็นประโยคที่ไม่ถูกต้องตามไวยากรณ์ เนื่องจากโดยปกติแล้วมนุษย์เราจะพูดแต่ภาษาที่ถูกต้องและสื่อความได้ เช่น เราคงจะไม่เคยได้ยินประโยคว่า "วิธีสิ่งสรรพอย่างทุกอยู่มีที่คิดความ" หรือประโยคบางประโยคก็ยากที่จะพบเห็นในชีวิตจริง แต่คนเราก็สามารถตัดสินได้ว่าประโยคนั้นเป็นประโยคที่ถูกไวยากรณ์หรือไม่ (grammatical) ดังเช่น ประโยคที่ Chomsky ยกขึ้นมาเป็นตัวอย่างคือ “colorless green idea sleeps furiously” (Chomsky 1957) ซึ่ง Chomsky กล่าวว่าประโยคลักษณะนี้เราสามารถระบุได้ว่าเป็นประโยคที่ถูกไวยากรณ์ถึงแม้ว่าจะไม่มีผู้ใดใช้หรือพูดประโยคนี้ในชีวิตประจำวันก็ตาม

---

[1] หมายถึงประโยคที่ผู้พูดที่เป็นเจ้าของภาษานั้นจะไม่พูดหรือใช้ เนื่องจากไม่มีกฎเกณฑ์ใดใดทางภาษาที่จะทำให้สร้างประโยคนั้นออกมาได้ ซึ่งจะต่างจากประโยคที่ผิดไวยากรณ์ตามความเข้าใจแบบทั่วไปที่อาจหมายรวมถึงประโยคที่อาจารย์ผู้สอนภาษาเห็นว่าใช้ไม่ถูกต้องหรือไม่เหมาะสม แต่แท้จริงแล้วเป็นประโยคที่ถูกไวยากรณ์เพราะมีโครงสร้างประโยคที่เป็นไปตามกฎไวยากรณ์ได้

3. เป็นไปได้ยากที่จะศึกษาคลังข้อมูลภาษาที่มีขนาดใหญ่ในลักษณะเชิงประจักษ์ (คำกล่าวนี้ อาจเป็นจริงในยุคนั้นซึ่งคอมพิวเตอร์ยังไม่มีประสิทธิภาพพอและยังไม่เป็นที่ใช้กันแพร่หลาย แต่ปัจจุบันคอมพิวเตอร์มีประสิทธิภาพมากพอที่จะใช้เก็บและค้นหาข้อมูลขนาดใหญ่ได้ คำกล่าวข้อนี้จึงไม่มีน้ำหนักพอ)

ถึงแม้ว่าการศึกษาภาษาศาสตร์โดยใช้ข้อมูลภาษาจริงจะลดบทบาทความสำคัญลงไปในระยะนั้น แต่ก็ยังมีนักภาษาศาสตร์จำนวนหนึ่งที่ยังเห็นถึงความสำคัญของคลังข้อมูลภาษาต่อการศึกษาภาษา อีกทั้งงานบางอย่างเองก็หลีกเลี่ยงไม่ได้ที่จะต้องอาศัยข้อมูลภาษาจริงเป็นพื้นฐานในการศึกษา เช่น การรับภาษา (language acquisition), ภาษาศาสตร์เชิงประวัติ (historical linguistics), ภาษาศาสตร์ภาคสนาม (linguistic field method) ในยุคนี้ จึงมีผู้เริ่มสร้างคลังข้อมูลภาษาต่างๆ เพื่อใช้ประโยชน์ในงานของตน ตัวอย่างเช่น

ในปี ค.ศ. 1960 Quirk ได้สร้างคลังข้อมูล “Survery of English Usage” (SEU) เป็นคลังข้อมูลภาษาอังกฤษแบบอังกฤษ (British English) ซึ่งรวบรวมทั้งภาษาพูดและภาษาเขียนที่ใช้ในระหว่างปี 1953-1987 คลังข้อมูลมีขนาดหนึ่งล้านคำโดยแบ่งเป็นข้อมูลภาษาพูดและภาษาเขียนอย่างละครึ่งหนึ่ง

ในปี ค.ศ. 1961 Francis และ Kucera สร้างคลังข้อมูล Brown Corpus เป็นคลังข้อมูลภาษาอังกฤษแบบอเมริกัน (American English) ที่เก็บจากเอกสารสิ่งพิมพ์ทั่วไปในปีนั้น คลังข้อมูลมีขนาดหนึ่งล้านคำ

ในปี ค.ศ.1975 Svartvik สร้างคลังข้อมูล London-Lund corpus (LLC) เป็นคลังข้อมูลภาษาอังกฤษแบบอังกฤษเฉพาะที่เป็นภาษาพูดจำนวน 500,000 คำ โดยนำข้อมูลจากคลังข้อมูล SEU มาบันทึกลงในระบบคอมพิวเตอร์

คลังข้อมูลภาษาในยุคต่อมาจะมีขนาดใหญ่ขึ้น เช่น Birmingham Collection of English Language Corpus ของ Sinclair มีขนาด 20 ล้านคำ

เป็นข้อมูลภาษาอังกฤษแบบอังกฤษ และในปัจจุบันคลังข้อมูลภาษาโดยทั่วไปก็จะมีขนาดใหญ่ตั้งแต่หนึ่งร้อยล้านคำขึ้นไป สามารถจัดเก็บและสืบค้นได้ในระบบคอมพิวเตอร์ ซึ่งก็เป็นสาเหตุหนึ่งที่ทำให้ภาษาศาสตร์คลังข้อมูล (corpus linguistics) เป็นที่สนใจแพร่หลาย เนื่องมาจากประสิทธิภาพของคอมพิวเตอร์ทั้งทางด้านความเร็วและความจุได้มีการพัฒนาขึ้นอย่างรวดเร็ว เราสามารถใช้คอมพิวเตอร์ในการค้นหาคำวลีหรือข้อความในคลังข้อมูลภาษาขนาดใหญ่ได้อย่างแม่นยำและรวดเร็ว สามารถนับความถี่ คำนวณค่าทางสถิติ และจัดเรียงข้อมูลในรูปแบบที่ต้องการได้ ทำให้การศึกษาแง่มุมต่างๆของภาษาเป็นไปได้โดยสะดวกมากขึ้น และทำให้เห็นรูปแบบการใช้ในลักษณะต่างๆที่บางครั้งก็ยากที่จะนึกถึงได้เอง

นับจากทศวรรษที่ 1980 มา นักภาษาศาสตร์จึงเริ่มให้ความสนใจกับข้อมูลภาษาจริงมากขึ้น คำว่า corpus linguistics หรือภาษาศาสตร์คลังข้อมูลจึงเป็นคำที่เกิดขึ้นมาในช่วงนี้ อาจจะด้วยเหตุผลที่ต้องการแสดงจุดยืนที่ต่างจากนักภาษาศาสตร์กระแสหลักที่ศึกษาภาษาโดยใช้ความหยั่งรู้ (intuition) เป็นหลัก ในปัจจุบัน การศึกษาทางภาษาศาสตร์เองก็ได้หันกลับมานิยมการศึกษาเชิงประจักษ์มากขึ้นโดยมีการใช้คลังข้อมูลภาษาเป็นพื้นฐานสำหรับการศึกษา โดยมีผู้ให้เหตุผลเพื่อสนับสนุนและชี้ให้เห็นข้อดีของการใช้ข้อมูลภาษาจริง เมื่อเทียบกับการศึกษาภาษาศาสตร์แบบเชิงเหตุผลซึ่งมักอาศัยข้อมูลภาษาที่คิดสังเคราะห์ขึ้นเองเป็นหลัก ดังเช่น

1. ข้อมูลสังเคราะห์ที่นักภาษาศาสตร์สร้างขึ้นเองเพื่อใช้ในการศึกษานั้นตรวจสอบหรือตัดสินความถูกต้องของการวิเคราะห์ได้ยาก ขึ้นอยู่กับการตัดสินของแต่ละคน ในขณะที่ข้อมูลจริงสามารถตรวจสอบได้ว่ามีการใช้แบบนั้นๆจริงไหม อีกทั้งข้อมูล

สังเคราะห์ที่นักภาษาศาสตร์ใช้กันบางครั้งฟังดูไม่เป็นธรรมชาติ และแทบจะไม่พบในบริบทจริงเลย ทำให้มีความเห็นแตกต่างกันได้

2. แนวทางของนักภาษาศาสตร์ในการแยกข้อมูลภาษาว่าเป็นประโยคที่ถูกไวยากรณ์หรือเป็นประโยคที่ผิดไวยากรณ์นั้นไม่เพียงพอที่จะอธิบายธรรมชาติที่แท้จริงของภาษา ในความเป็นจริงมีประโยคจำนวนไม่น้อยที่ตัดสินไม่ได้ง่ายว่าถูกหรือผิดไวยากรณ์ แต่อาจเป็นที่ยอมรับได้ของเจ้าของภาษาในระดับที่แตกต่างกัน

3. คลังข้อมูลภาษาช่วยให้เราทราบถึงความถี่ของการใช้คำหรือโครงสร้างประโยคแบบต่างๆ ในขณะที่การศึกษาโดยข้อมูลสังเคราะห์ไม่อาจชี้ให้เห็นข้อมูลนี้ได้ และคลังข้อมูลภาษายังช่วยให้เราเห็นตัวอย่างหรือ ปรากฏการณ์ทางภาษาที่บางครั้งก็ยากที่จะนึกถึงได้โดยไม่ศึกษาจากข้อมูลจริง

ภาษาศาสตร์คลังข้อมูลจึงอาจมองได้ว่าเป็นเพียงแนวการศึกษาทางภาษาศาสตร์ที่อาศัยข้อมูลภาษาจริงเป็นพื้นฐานในการวิจัย ซึ่งอาจเป็นการศึกษาวิจัยในแขนงใดใดก็ได้ เช่น สัทศาสตร์ วากยสัมพันธ์ ปริจเฉทวิเคราะห์ ฯลฯ ภาษาศาสตร์คลังข้อมูลจึงไม่ใช่แขนงของภาษาศาสตร์ในความหมายที่เป็นแขนงหนึ่งที่มีขอบเขตการศึกษาภาษาที่ชัดเจน อย่างไรก็ตาม หากมองว่าภาษาศาสตร์คลังข้อมูลนอกจากจะสนใจตอบสมมติฐานต่างๆโดยอาศัยข้อมูลภาษาแล้ว ยังสนใจเรื่องของการจัดสร้างคลังข้อมูลภาษาที่จะเป็นประโยชน์ต่อการศึกษาภาษาในเรื่องอื่นๆ ต่อไป และสนใจเรื่องของเครื่องมือที่ใช้ในการทำงานร่วมกับคลังข้อมูลภาษา ภาษาศาสตร์คลังข้อมูลก็อาจเป็นแขนงหนึ่งทางภาษาศาสตร์ที่แตกต่างจากภาษาศาสตร์แขนงอื่น ๆ ในปัจจุบัน ภาษาศาสตร์คลังข้อมูลเป็นแขนงวิชาที่ได้รับความสนใจเป็นอย่างมาก เพราะนอกจากประสิทธิภาพของคอมพิวเตอร์จะช่วยให้การจัดสร้างและใช้ประโยชน์ในการวิจัยภาษาศาสตร์เป็นไปได้โดยสะดวกแล้ว คลังข้อมูลภาษายังเป็นประโยชน์ในด้าน

อื่นๆ อีกด้วย เช่น ช่วยในการจัดทำพจนานุกรมและประมวลศัพท์เฉพาะสาขา เพราะทำให้ผู้จัดทำสามารถเขียนคำนิยามศัพท์โดยการสืบค้นความหมายจากข้อมูลจริงได้ ช่วยในการเรียนการสอนภาษา เพราะทำให้นักเรียนเรียนรู้การใช้ภาษาที่เกิดขึ้นจริงได้ นอกจากนี้คลังข้อมูลภาษายังเป็นประโยชน์ในงานด้านการประมวลผลภาษาธรรมชาติ (natural language processing) คลังข้อมูลภาษาสามารถนำมาใช้เพื่อฝึกสอน (train) คอมพิวเตอร์ให้รู้ถึงลักษณะบางประการของภาษาเพื่อให้สามารถประมวลผลภาษาได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งในปัจจุบันนี้ สามารถกล่าวได้ว่าการใช้คลังข้อมูลภาษาเป็นกระแสหลักของภาษาศาสตร์คอมพิวเตอร์

3. ขอบเขตของภาษาศาสตร์คลังข้อมูล

ภาษาศาสตร์คลังข้อมูลไม่ใช่แขนงทางภาษาศาสตร์ที่ศึกษาทฤษฎีทางภาษาศาสตร์ในลักษณะเดียวกับทฤษฎีไวยากรณ์ต่างๆ ไม่ได้เป็นแขนงทางภาษาศาสตร์ที่มีขอบเขตการศึกษาภาษากับเรื่องใดเรื่องหนึ่งเหมือนอย่างภาษาศาสตร์สังคม ภาษาศาสตร์จิตวิทยา หรือภาษาศาสตร์คอมพิวเตอร์ แต่ภาษาศาสตร์คลังข้อมูลเป็นแขนงที่สนใจเรื่องการจัดสร้างคลังข้อมูลภาษาและเครื่องมือที่จะใช้งานร่วมกับคลังข้อมูลภาษา และการใช้ประโยชน์จากคลังข้อมูลภาษาเพื่อศึกษาวิจัยประเด็นต่างๆที่เกี่ยวข้องกับภาษา

โดยปกติ ในการศึกษาทางภาษาศาสตร์นั้นจำเป็นต้องอาศัยหลักฐาน คือข้อมูลภาษามาเพื่อยืนยันทฤษฎีที่ใช้อธิบายภาษาดังกล่าว ซึ่งหลักฐานดังกล่าวนี้บางครั้งก็ได้จากการนึกคิดเอาโดยอาศัยความหยั่งรู้ของเจ้าของภาษาในการตัดสินว่าประโยคนั้นๆถูกไวยากรณ์หรือไม่ สำหรับการศึกษาแบบภาษาศาสตร์คลังข้อมูลนั้น การนำคลังข้อมูลภาษามาใช้จะช่วยให้ได้หลักฐานข้อมูลการใช้ภาษาที่เกิดขึ้นจริง อันจะนำไปสู่ผลการศึกษาที่ดีกว่าการศึกษาโดย

ใช้ข้อมูลในลักษณะอื่น ๆ การใช้ประโยชน์จากคลังข้อมูลภาษาอย่างเต็มที่นั้นจำเป็นต้องมีโปรแกรมคอมพิวเตอร์เพื่อช่วยในการสืบค้นข้อมูลภาษาในลักษณะต่างๆ และต้องมีการออกแบบจัดเก็บข้อมูลภาษาอย่างเป็นระบบ งานทางด้านภาษาศาสตร์คลังข้อมูลจึงเกี่ยวข้องกับกลุ่มบุคคลหลายกลุ่ม กลุ่มแรกเป็นผู้ที่เกี่ยวข้องกับการออกแบบและสร้างคลังข้อมูลภาษา (ดูบทที่ 8-10) กลุ่มที่สองเป็นผู้ที่เกี่ยวข้องกับการสร้างโปรแกรมคอมพิวเตอร์ที่จะใช้ร่วมกับคลังข้อมูลภาษา (ดูบทที่ 11 และ 12) กลุ่มที่สามเป็นผู้ที่ใช้ประโยชน์จากคลังข้อมูลภาษาเพื่องานของตน (ดูบทที่ 2-7) เช่น นักภาษาศาสตร์ที่ใช้คลังข้อมูลภาษาเพื่อศึกษาถึงลักษณะการใช้ภาษาที่เกิดขึ้นจริง ความถี่ในการปรากฏของคำหรือวลีที่ศึกษา อาจารย์สอนภาษาที่ใช้คลังข้อมูลภาษาเพื่อการวิจัยและการเรียนการสอนภาษา นักแปลที่อาศัยคลังข้อมูลภาษาหาตัวอย่างของการแปล นักทฤษฎีการแปลที่ศึกษาธรรมชาติของการแปลจากคลังข้อมูลเทียบบท (parallel corpora) และนักภาษาศาสตร์คอมพิวเตอร์ที่ใช้คลังข้อมูลภาษาเพื่อช่วยในการพัฒนาระบบการประมวลผลภาษาธรรมชาติ เป็นต้น

4. พัฒนาการของคลังข้อมูลภาษา[2]

การใช้ประโยชน์จากคลังข้อมูลภาษานั้น เริ่มมีมานานก่อนที่จะมีการใช้คอมพิวเตอร์กันอย่างแพร่หลาย ในอเมริกาและยุโรปในช่วงทศวรรษที่ 1920 ได้มีการนับความถี่ของคำที่ใช้กันในภาษาอังกฤษเพื่อช่วยในการจัดทำหลักสูตรการสอนที่เหมาะสมกับการเรียนการสอนภาษาอังกฤษ ตัวอย่างเช่น ในปีค.ศ. 1921 Thorndike ได้รวบรวมข้อมูลภาษาอังกฤษขนาด 4.5 ล้านคำเพื่อสร้างรายการ

---

[2] เนื้อหาส่วนใหญ่ในหัวข้อนี้เรียบเรียงจากบทที่ 2 ในหนังสือของ Kennedy (1998)

ความถี่ของคำ (word frequency list) เพื่อที่จะใช้เป็นแนวทางในการเตรียมเนื้อหาการสอน (material) ที่เหมาะสมต่อการสอนวิชาวรรณคดี สำหรับนักเรียนในอเมริกาที่พูดภาษาอังกฤษเป็นภาษาแม่ ต่อมา Thorndike ได้ร่วมมือกับ Lorge จัดทำคลังข้อมูลที่มีขนาดใหญ่ขึ้นคือมีขนาด 18 ล้านคำในปีค.ศ. 1944 ผลงานของ Thorndike และ Lorge ชิ้นนี้มีผลอย่างมากต่อการจัดการเรียนการสอนภาษาอังกฤษในประเทศต่างๆทั่วโลกในเวลาต่อมา

ในงานด้านการเขียนตำราไวยากรณ์ ก็มีผู้ใช้ประโยชน์จากคลังข้อมูลภาษาเช่นกัน ตัวอย่างเช่น ในปีค.ศ. 1940 Fries ใช้จดหมายที่บุคคลต่างๆเขียนถึงรัฐบาลสหรัฐอเมริกาเพื่อเป็นข้อมูลในการเขียนตำรา “American English Grammar” และใช้บทสนทนาทางโทรศัพท์ขนาด 250,000 คำเป็นข้อมูลสำหรับเขียนหนังสือ “The Structure of English” อย่างไรก็ตามงานที่สำคัญที่สุดในยุคนี้ คือ การจัดสร้างคลังข้อมูล Survey of English Usage (SEU) ของ Quirk ซึ่งเริ่มต้นในปีค.ศ. 1958 โดยตั้งเป้าที่จะรวบรวมตัวอย่างของการใช้ภาษาอังกฤษของผู้มีการศึกษาดีจำนวน 200 ตัวอย่าง แต่ละตัวอย่างมีขนาด 5,000 คำ เพื่อใช้เป็นตัวแทนของภาษาอังกฤษแบบอังกฤษ (British English) คลังข้อมูลมีขนาดรวมหนึ่งล้านคำ เป็นข้อมูลภาษาพูดและภาษาเขียนอย่างละครึ่งหนึ่ง โดยตัวบท (text) ที่รวบรวมไว้ใน SEU เป็นเอกสารภาษาที่อยู่ในช่วงระหว่างปี ค.ศ. 1953-1987 โครงการนี้ใช้เวลาดำเนินการทั้งสิ้นกว่า 25 ปี โดยเริ่มตั้งแต่สมัยที่ยังไม่มีการใช้คอมพิวเตอร์จนกระทั่งมาเสร็จสมบูรณ์ในปีค.ศ. 1989 ซึ่งในภายหลังข้อมูลในส่วนของภาษาพูดได้ถูกนำมาแยกจัดเก็บในคอมพิวเตอร์เป็น London-Lund Corpus

ส่วนในงานด้านการทำพจนานุกรม (lexicography) และการศึกษาภาษาถิ่น (dialectology) นั้นจำเป็นอย่างยิ่งที่ต้องอาศัยข้อมูลภาษาจริงเป็นหลัก

ตัวอย่างของการศึกษาทางด้านนี้ ได้แก่ พจนานุกรม Oxford English Dictionary ที่ตีพิมพ์ในปีค.ศ. 1928 เป็นผลงานที่ได้มาจากการศึกษาข้อมูลภาษาจริงอย่างต่อเนื่องมาเป็นเวลากว่า 71 ปี ใช้คนกว่า 2,000 คนในการรวบรวมประโยคอ้างอิง (citation) 5 ล้านประโยคหรือกว่า 50 ล้านคำสำหรับใช้ประกอบการเขียนคำอธิบายศัพท์ 414,825 ศัพท์ และพจนานุกรม Webster New International Dictionary ที่ตีพิมพ์ในปีค.ศ. 1961 เป็นพจนานุกรมฉบับสุดท้ายที่ใช้คลังข้อมูลภาษาที่ไม่ได้เก็บอยู่ในระบบคอมพิวเตอร์ พจนานุกรมนี้มีประโยคอ้างอิงกว่า 10 ล้านประโยคที่ใช้ประกอบคำอธิบายศัพท์ หลังจากนั้น จึงเริ่มมีการนำคอมพิวเตอร์เข้ามาใช้ ซึ่งในยุคแรกยังเป็นเครื่องระดับเมนเฟรม และใช้ช่วยในการวิเคราะห์ข้อมูลระดับคำและคำที่ปรากฏร่วมกัน (word collocation) เป็นหลัก

4.1 คลังข้อมูลยุคแรก

เมื่อคลังข้อมูลภาษาถูกจัดเก็บไว้ในคอมพิวเตอร์ งานเกี่ยวกับการค้นหาคำ นับความถี่ของคำ ก็สามารถทำได้สะดวกเป็นอย่างมาก ดังนั้นงานต่างๆที่เกี่ยวข้องกับภาษา เช่น การทำพจนานุกรม งานวิจัยทางภาษาศาสตร์ งานด้านการเรียนการสอนภาษา จึงเริ่มให้ความสนใจกับการใช้ การจัดสร้างคลังข้อมูล ซึ่งคลังข้อมูลภาษาในยุคแรกๆจะยังคงมีขนาดเล็กอยู่ คือ มีขนาดประมาณหนึ่งล้านคำ ได้แก่

4.1.1 Brown Corpus

คลังข้อมูล Brown เป็นคลังข้อมูลขนาดหนึ่งล้านคำของภาษาเขียนในภาษาอังกฤษแบบอเมริกัน ซึ่งเริ่มดำเนินการสร้างเมื่อปีค.ศ. 1961 และเสร็จสิ้นในปีค.ศ. 1964 โครงสร้างคลังข้อมูลประกอบด้วยตัวบท 500 ตัวอย่าง แต่ละตัวอย่างมีขนาด 2,000 คำ โดยสุ่มเลือกจากเอกสารประเภทต่างๆ เป็นตัวบทที่

ให้ข้อมูลข่าวสาร (informative) ร้อยละ 75 ตัวบทที่เป็นเรื่องจินตนาการ (imaginative prose) ร้อยละ 25 ซึ่งในตัวบทแต่ละประเภทก็จะแบ่งเป็นหมวดย่อยๆ เพื่อให้ครอบคลุมภาษาเขียนแบบต่างๆให้มากที่สุด เสมือนให้เป็นตัวแทน (representative) ของภาษาเขียน โครงสร้างของคลังข้อมูล Brown จะคล้ายกับโครงสร้างของคลังข้อมูล SEU เนื่องจาก ผู้สร้างคลังข้อมูล Brown คือ Francis และ Kucera ได้รับอิทธิพลมาจาก Quirk ซึ่งเป็นผู้กำหนดโครงสร้างของคลังข้อมูล SEU

คลังข้อมูล Brown เริ่มสร้างโดยการพิมพ์บัตรเจาะรูคอมพิวเตอร์แล้วถ่ายเข้าเทปแม่เหล็กในระบบคอมพิวเตอร์เมนเฟรม แต่ปัจจุบันจัดเก็บและเผยแพร่ในรูปของซีดีรอม คลังข้อมูล Brown เป็นต้นแบบของคลังข้อมูลภาษาที่ออกแบบมาเพื่อให้เสมือนเป็นตัวแทนของภาษา (representative corpus) ที่มีการสร้างกันต่อมา

4.1.2 Lancaster-Oslo/Bergen (LOB) Corpus

คลังข้อมูล Lancaster-Oslo/Bergen (LOB) เป็นคลังข้อมูลคู่เทียบของคลังข้อมูล Brown โดยคลังข้อมูล LOB เป็นคลังข้อมูลภาษาเขียนของภาษาอังกฤษแบบอังกฤษ สร้างขึ้นในช่วงปีค.ศ. 1970-1978 โดยความร่วมมือของมหาวิทยาลัย Lancaster มหาวิทยาลัย Oslo และ Norwegian Computing Centre for the Humanities at Bergen คลังข้อมูล LOB ประกอบด้วยตัวบท 500 ตัวอย่าง แต่ละตัวอย่างมีขนาด 2,000 คำ การจัดหมวดเอกสารใน LOB นั้นจะเหมือนกับคลังข้อมูล Brown แต่อาจมีความแตกต่างกันบ้างเล็กน้อย เช่น ในคลังข้อมูล LOB จะมีจำนวนตัวบทในส่วนที่เป็นเรื่องแบบลูกทุ่งตะวันตก (western) น้อยกว่า แต่จะเพิ่มส่วนที่เป็นการผจญภัย (adventure) เข้ามาแทนเป็นต้น

คลังข้อมูลที่มีโครงสร้างเหมือนกับคลังข้อมูล Brown นอกจากคลังข้อมูล LOB แล้วยังมีคลังข้อมูลอื่นๆ อีก เช่น Kolhapur Corpus of Indian English, Wellinton Corpus of Written New Zealand English, และ Australian Corpus of English (ACE) ซึ่งต่างก็เป็นคลังข้อมูลของภาษาอังกฤษที่ใช้ในประเทศนั้นๆ

4.1.3 London-Lund Corpus (LLC)

คลังข้อมูล London-Lund เป็นคลังข้อมูลส่วนที่เป็นภาษาพูดของคลังข้อมูล SEU คลังข้อมูล SEU นั้นเริ่มแรกไม่ได้ถูกสร้างมาในรูปของข้อความอิเล็กทรอนิกส์ ข้อมูลทั้งภาษาพูดและภาษาเขียนถูกวิเคราะห์และเก็บลงบนกระดาษ จนกระทั่งในปีค.ศ. 1975 Jan Svartvik แห่งมหาวิทยาลัย Lund ในประเทศสวีเดน ได้เริ่มโครงการ Survey of Spoken English (SSE) โดยตั้งเป้าหมายว่าจะจัดเก็บข้อมูลภาษาพูดในรูปของข้อความอิเล็กทรอนิกส์ โดยใช้ข้อมูลที่ได้มาจากคลังข้อมูล SEU ที่เป็นตัวบทภาษาพูดจำนวน 87 ตัวบท และเพิ่มเติมตัวบทอีก 13 ชิ้นเพื่อให้ครบหนึ่งร้อยชิ้น แต่ละตัวบทมีขนาด 5,000 คำ รวมทั้งหมดเป็นคลังข้อมูลภาษาขนาดหนึ่งล้านคำ คลังข้อมูล LLC จัดว่าเป็นคลังข้อมูลภาษาพูดที่มีการถอดลักษณะต่างๆของเสียงไว้อย่างละเอียด แต่คลังข้อมูล LLC ก็มีลักษณะที่เน้นเนื้อหาทางวิชาการมาก เพราะข้อมูลส่วนใหญ่ได้จากกลุ่มคนที่มีการศึกษา ซึ่งทำงานทางด้านวิชาการ นอกจากนี้การจัดประเภทหมวดเนื้อหาต่างๆในคลังข้อมูลก็มีความหลากหลายน้อยกว่าเมื่อเทียบกับ International Corpus of English (ICE) (ดู 4.2.4)

4.2 คลังข้อมูลยุคที่สอง

คลังข้อมูลภาษาขนาดหนึ่งล้านคำในยุคแรกนั้นแม้จะจัดได้ว่ามีขนาดใหญ่ แต่ก็ยังเป็นคลังข้อมูลที่มีขนาดจำกัดเกินกว่าที่จะใช้วิเคราะห์ศัพท์ (lexical analysis) หรือใช้วิเคราะห์เรื่องความหมาย (semantic analysis) ได้ เนื่องจากคำบางคำอาจมีความถี่ในการปรากฏน้อยมากเพียงหนึ่งครั้งหรืออาจไม่ปรากฏเลยในคลังข้อมูลขนาดหนึ่งล้านคำนี้ ดังนั้น ในบางครั้ง จึงจำเป็นต้องอาศัยคลังข้อมูลที่มีขนาดใหญ่มากๆ เพื่อให้มีข้อมูลคำนั้นในปริมาณที่มากเพียงพอสำหรับการวิเคราะห์ ซึ่งในปัจจุบันตั้งแต่ปี ค.ศ. 1990 เป็นต้นมา มาตรฐานของขนาดคลังข้อมูลภาษาจะอยู่ที่ประมาณ 100 ล้านคำ

4.2.1 The Cobuild Project

โครงการ Cobuild เป็นความร่วมมือของหลายหน่วยงาน คือ สำนักพิมพ์ Collins และทีมวิจัยในมหาวิทยาลัย Birmingham คลังข้อมูลที่สร้างขึ้นมานี้ มีชื่อว่า Cobuild Corpus หรือ Birmingham Corpus โครงสร้างข้อมูลของคลังข้อมูล Cobuild ในส่วนที่เป็นคลังข้อมูลหลัก (core corpus) ถูกออกแบบมาเพื่อให้เป็นตัวแทน (representative) ของภาษาอังกฤษ เพื่อให้เป็นประโยชน์แก่ผู้เรียน ผู้สอนภาษา และผู้สนใจภาษาในด้านอื่นๆ โดยกำหนดให้ร้อยละ 25 ของคลังข้อมูลเป็นภาษาพูด ข้อมูลทั้งหมดรวบรวมจากเอกสารที่ปรากฏหลังปี ค.ศ. 1960 ไปแล้ว ซึ่งถือว่าเป็นภาษาปัจจุบัน และจะเน้นภาษาที่เป็นภาษาอังกฤษทั่วไปมากกว่าภาษาเฉพาะทาง (technical text) โดยมีสัดส่วนเป็นภาษาอังกฤษแบบอังกฤษ (British English) ร้อยละ 70 ภาษาอังกฤษแบบอเมริกัน (American English) ร้อยละ 20 และภาษาอังกฤษแบบอื่นๆอีกร้อยละ 10 ลักษณะของข้อมูลภาษาพูดจะเน้นภาษาที่เป็นภาษาพูดจริงๆ มากกว่าภาษาพูดที่ปรากฏในบทละคร และมีการกำหนดอายุของผู้พูด คือต้องเป็นบุคคลที่มีอายุ 16 ปีขึ้นไป

สำหรับข้อมูลภาษาเขียนจะมีการกำหนดเกณฑ์ต่างๆเพื่อใช้คัดเลือกตัวบท เช่น ต้องเป็นเรื่องที่มีผู้นิยมอ่านมาก และแบ่งเป็นตัวบทที่ผู้แต่งเป็นเพศชายร้อยละ 75 และตัวบทที่ผู้แต่งเป็นเพศหญิงร้อยละ 25 เป็นต้น

ในระยะแรกของการดำเนินการสร้างคลังข้อมูล คือในปี ค.ศ. 1982 คลังข้อมูลหลักมีขนาด 7.3 ล้านคำ ต่อมาหลังจากตีพิมพ์ Collins Cobuild English Dictionary ในปี 1987 แล้ว ก็ได้มีการเพิ่มจำนวนข้อมูลให้มากขึ้นโดยไม่เน้นการเป็นคลังข้อมูลที่สมดุล (balance corpus) อีกต่อไป ข้อมูลในส่วนที่เพิ่มนี้เรียกคลังข้อมูลเสริม (reserve corpus) มีขนาด 13 ล้านคำ ต่อมาในปี ค.ศ. 1990 John Sinclair ผู้อำนวยการของโครงการ Cobuild ได้ดำเนินการขยายโครงการให้เป็นคลังข้อมูลภาษาอังกฤษขนาดใหญ่เรียกว่า The Bank of English โดยเป็นคลังข้อมูลขนาดมากกว่าหนึ่งร้อยล้านคำ และสามารถขยายขนาดใหญ่ขึ้นได้ตามจำนวนข้อมูลที่เพิ่มมากขึ้น ซึ่งก็ปรากฏว่าในปี ค.ศ. 2009 คลังข้อมูลนี้ได้มีขนาดถึง 524 ล้านคำ[3]

4.2.2 The Longman Corpus Network

เป็นฐานข้อมูลเชิงพาณิชย์ซึ่งประกอบด้วยคลังข้อมูลภาษาอังกฤษสามคลังข้อมูล คือ Longman/Lancaster English Language Corpus (LLELC), Longman Spoken Corpus (LSC) และ Longman Corpus of Learners' English (LCLE) แต่ละคลังข้อมูลมีโครงสร้างและเนื้อหาข้อมูลที่แตกต่างกัน และถูกจัดสร้างขึ้นมาเพื่อวัตถุประสงค์ที่แตกต่างกัน คลังข้อมูลทั้งสามนี้ใช้เพื่อประโยชน์ในการทำพจนานุกรม

[3] ข้อมูลจาก http://www.collins.co.uk/books.aspx?group=153

คลังข้อมูล LCLE มีขนาดประมาณสิบล้านคำและสร้างโดยการรวบรวมตัวอย่างภาษาเขียนภาษาอังกฤษจากจดหมาย รายงาน กระดาษคำตอบ บันทึกประจำวัน ข้อเขียนของนักเรียนที่คัดจากนักเรียนในระดับต่างๆแปดระดับจากผู้เรียนภาษาอังกฤษที่พูดภาษาแม่แบบต่าง ๆ 160 ภาษา ส่วนคลังข้อมูล LSC นั้นต่อมาได้กลายมาเป็นส่วนหนึ่งของ British National Corpus ส่วนคลังข้อมูล LLELC นั้น เริ่มต้นสร้างในปลายทศวรรษที่ 80 โดยทีมวิจัยของ Della Summers โดยตั้งเป้าให้เป็นคลังข้อมูลภาษาแบบสมดุลของภาษาอังกฤษในศตวรรษที่ 20 คลอบคลุมทั้งภาษาอังกฤษแบบอังกฤษ ภาษาอังกฤษแบบอเมริกันและภาษาอังกฤษแบบอื่นๆ และต้องการให้มีขนาดถึง 50 ล้านคำ วิธีที่ใช้ในการออกแแบบให้เป็นคลังข้อมูลที่เป็นเสมือนตัวแทนภาษา (representative corpus) ใช้สองวิธีผสมกัน คือ สุ่มเลือกหนังสือจากฐานข้อมูล Book In Print มาทำคลังข้อมูลกึ่งหนึ่ง ส่วนอีกกึ่งหนึ่งคัดเลือกโดยดูเกณฑ์ต่างๆตามที่กำหนด เช่น ความเป็นที่นิยมของเอกสาร ความสำคัญของเอกสาร เป็นต้น และเพื่อให้คลังข้อมูลมีน้ำหนักที่สมดุล จึงควบคุมโดยกำหนดให้คลังข้อมูลประกอบด้วยตัวบทที่เป็นข้อมูลข่าวสาร (informative) ร้อยละ 58 ตัวบทประเภทจินตนาการ (imaginative) ร้อยละ 42 โดยที่ตัวบทที่ใช้นั้นคัดจากหนังสือกว่า 2,000 เล่ม แต่ละเล่มคัดเลือกมา 40,000 คำ

4.2.3 The British National Corpus

โครงการคลังข้อมูล British National Corpus หรือ BNC เริ่มดำเนินการในปี ค.ศ. 1991 โดยความร่วมมือของหลายๆหน่วยงาน ทั้งจากสำนักพิมพ์และมหาวิทยาลัยต่างๆ ได้แก่ สำนักพิมพ์ Oxford University Press, Longman Group, W & R Chambers, British Library, มหาวิทยาลัย Oxford และมหาวิทยาลัย Lancaster โดยค่าใช้จ่ายครึ่งหนึ่งของโครงการได้รับการสับสนุน

จากรัฐบาลอังกฤษ โครงการนี้มีวัตถุประสงค์เพื่อสร้างคลังข้อมูลภาษาขนาด 100 ล้านคำ โดยรวบรวมภาษาพูดและภาษาเขียนในภาษาอังกฤษแบบอังกฤษในปัจจุบัน (British English) คลังข้อมูลนี้ได้รับการออกแบบเพื่อให้เป็นตัวแทน (representative) ของภาษาอังกฤษแบบอังกฤษ โดยไม่จำกัดเฉพาะภาษาอังกฤษถิ่นใดถิ่นหนึ่ง แบ่งเป็นข้อมูลภาษาเขียนร้อยละ 90 และข้อมูลภาษาพูดร้อยละ 10 โครงสร้างข้อมูลภาษาเขียนประกอบด้วยเนื้อหาที่เป็นข้อมูลข่าวสารร้อยละ 75 และเนื้อหาที่เป็นจินตนาการร้อยละ 25 ที่มาของข้อมูลภาษาเขียนทั้งหมดได้มาจากหนังสือร้อยละ 60 ได้มาจากวารสารร้อยละ 25 ได้มาจากเอกสารแผ่นพับเพื่อการเผยแพร่ (brochure) ร้อยละ 5 และได้จากเอกสารที่ไม่ตีพิมพ์เผยแพร่อีกร้อยละ 5 นอกจากนั้น บางส่วนยังได้มาจากบทละคร สุนทรพจน์ หรือ บทที่เขียนขึ้นสำหรับพูด ในแง่การจัดตามระดับภาษา แบ่งเป็นภาษาระดับสูง เช่น วรรณกรรม ร้อยละ 30 ภาษาระดับกลางและระดับล่างร้อยละ 45 และ 25 ตามลำดับ

สำหรับส่วนที่เป็นข้อมูลภาษาพูดนั้น จะแบ่งเป็นสองส่วน ส่วนแรกได้แก่ ข้อมูลภาษาที่อิงกับบริบทการใช้ (context-governed) เช่น การสนทนาในชั้นเรียน การอบรม การรายงานข่าว การปราศรัยในที่สาธารณะ เป็นต้น ส่วนที่สองจะเป็นการบันทึกการสนทนาในชีวิตประจำวันของอาสาสมัครจำนวน 124 คน จากภูมิภาคต่างๆทั่วอังกฤษ รวมเป็นเวลาถึง 2,000 ชั่วโมง โดยแบ่งเป็นเพศชายและหญิงในจำนวนเท่าๆกัน ในช่วงอายุตั้งแต่ 15 ปี ถึงประมาณ 60 ปี โดยกำหนดให้อาสาสมัครติดเครื่องบันทึกเสียงซึ่งจะบันทึกการพูดคุยในชีวิตประจำวันเป็นเวลาสองวัน

การออกแบบคลังข้อมูลให้เป็นตัวแทนของภาษานับเป็นสิ่งสำคัญที่ต้องมีการพิจารณาอย่างละเอียดและรอบคอบ เนื่องจากการพยายามเข้าถึงรูปแบบ

ภาษาให้ครบถ้วนในทุกๆ ด้านนั้นทำได้ยาก ซึ่งมีผลทำให้การกำหนดเกณฑ์ที่เหมาะสมและถูกต้องในการคัดเลือกเนื้อหาแต่ละประเภทเป็นไปได้ยากด้วย ผู้จัดทำคลังข้อมูล BNC จึงแก้ปัญหานี้โดยพยายามทำให้ข้อมูลมีความหลากหลายมากที่สุดโดยใช้เกณฑ์คัดเลือกจากมุมมองในหลายมิติร่วมกัน ได้แก่ ในด้านการผลิต ซึ่งอาจพิจารณาจากรายการหนังสือที่มีการตีพิมพ์ ในด้านการบริโภค อาจพิจารณาจากรายการหนังสือที่มียอดขายสูง หรือมีจำนวนครั้งในการยืมจากห้องสมุดมากครั้ง เป็นต้น อนึ่ง ในการสร้างคลังข้อมูลภาษา BNC ได้มีการกำหนดเกณฑ์ 4 เกณฑ์เพื่อคัดเลือกข้อมูลภาษาจากตัวบทที่บรรจุเนื้อหาแต่ละประเภท ได้แก่ เกณฑ์ด้านเนื้อหา (domain) เกณฑ์ด้านเวลา (time) เกณฑ์ด้านสื่อที่ใช้ (medium) และเกณฑ์ด้านระดับภาษา (level) ตัวบทแต่ละส่วนที่นำมาใช้ในการสร้างข้อมูลภาษาจะถูกคัดเลือกมาจำนวน 40,000 คำ ซึ่งจะต้องถูกกำกับ (mark) ด้วยแท็ก (tag) ต่างๆตามมาตรฐาน SGML ต่อมาจึงมีการปรับมาตรฐานการกำกับข้อมูลโดยใช้ XML แทน (ดูบทที่ 10)

4.2.4 The International Corpus of English (ICE)

เป็นโครงการที่มีวัตถุประสงค์เพื่อรวบรวมข้อมูลภาษาอังกฤษแบบต่างๆ ทั่วโลก เริ่มดำเนินการในปี ค.ศ. 1988 โดยผู้อำนวยการของคลังข้อมูลภาษา SEU ได้เสนอจัดทำโครงการที่ใหญ่ขึ้นเพื่อการศึกษาเปรียบเทียบภาษาอังกฤษที่ใช้ในประเทศต่างๆ ในแต่ละภูมิภาคทั่วโลก ซึ่งได้แก่ประเทศที่ใช้ภาษาอังกฤษเป็นภาษาประจำชาติ หรือเป็นภาษาที่สอง เช่น อังกฤษ อเมริกา แคนาดา ออสเตรเลีย นิวซีแลนด์ อินเดีย ไนจีเรีย สิงคโปร์ เป็นต้น คลังข้อมูลที่สร้างขึ้นนี้จะประกอบด้วยคลังข้อมูลย่อยกว่า 20 คลังข้อมูล แต่ละคลังข้อมูลย่อยมีขนาดหนึ่งล้านคำ โดยรวบรวมภาษาที่ใช้โดยคนอายุ 18 ปีขึ้นไปและเป็นผู้มีการศึกษา ทั้งภาษาพูดและภาษาเขียนเลือกจากช่วงปี ค.ศ.1990-1993 โดยมีข้อมูลภาษาเขียน

ร้อยละ 40 ภาษาพูดร้อยละ 60 แต่ละคลังข้อมูลย่อยจะมีโครงสร้างแบบเดียวกัน คือประกอบไปด้วยตัวอย่างข้อความ 500 ตัวบท แต่ละตัวบทมีขนาด 2,000 คำ

5. ประเภทของคลังข้อมูลภาษา

การจำแนกประเภทของคลังข้อมูลภาษาสามารถจำแนกได้ตามเกณฑ์ต่างๆ ที่กำหนดขึ้น เช่น จำแนกตามวัตถุประสงค์ จำแนกตามลักษณะข้อมูล หรือจำแนกตามรูปแบบวิธีในการเก็บ หากเป็นการจำแนกตามวัตถุประสงค์การใช้ ก็แบ่งได้เป็นคลังข้อมูลภาษาแบบทั่วไป (general corpora) และคลังข้อมูลภาษาเฉพาะทาง (specialized corpora) ตัวอย่างของคลังข้อมูลภาษาแบบทั่วไป ได้แก่ คลังข้อมูล SEU คลังข้อมูลแบบนี้โดยทั่วไปจะมีลักษณะเป็นคลังข้อมูลแบบสมดุล (balance corpus) คือมีการเลือกเก็บข้อมูลภาษาอย่างมีระบบ โดยกำหนดสัดส่วนที่แน่นอนสำหรับการใช้แหล่งข้อมูลแต่ละแหล่ง ส่วนคลังข้อมูลภาษาเฉพาะทางซึ่งสร้างขึ้นเพื่อศึกษาภาษาเฉพาะทางจะเป็นคลังข้อมูลที่ผู้จัดทำคัดเลือกและรวบรวมเฉพาะข้อมูลของภาษาที่ต้องการศึกษาโดยเฉพาะ เช่น คลังข้อมูลภาษาเด็ก CHILDES (Child Language Data Excahange Systems) ของมหาวิทยาลัย Carnegie Mellon (http://childes.psy.cmu.edu/) เป็นต้น

หากพิจารณาจำแนกตามลักษณะข้อมูลภาษา ก็สามารถแบ่งได้เป็นคลังข้อมูลภาษาพูด หรือคลังข้อมูลภาษาเขียน คลังข้อมูลภาษาส่วนมากมักเป็นคลังข้อมูลภาษาเขียนเนื่องจากสามารถจัดสร้างหรือรวบรวมได้ง่ายกว่าภาษาพูด ซึ่งจะต้องมีการถอดเทปบทสนทนา และใส่ข้อมูลทางสัทลักษณ์ต่างๆ คลังข้อมูลบางอย่างก็สามารถเก็บข้อมูลตัวบททั้งหมดได้เพราะตัวบททั้งหมดมีขนาดที่จำกัดแน่นอน เช่น คลังข้อมูลงานเขียนของเช็คสเปียร์ แต่คลังข้อมูลทั่วไปมักเป็นการสุ่มเก็บตัวอย่างภาษา ซึ่งการสุ่มตัวอย่างก็อาจเป็นได้ทั้งการเก็บข้อความทั้งหมดของตัวอย่างนั้น หรือเก็บเฉพาะตัวอย่างในขนาดที่กำหนด เช่น 2,000 คำ

หากจำแนกตามลักษณะการจัดเก็บข้อมูล คลังข้อมูลภาษาก็แบ่งได้เป็นคลังข้อมูลที่เก็บเฉพาะตัวบทหรือข้อความล้วน (plain text) หรือเป็นคลังข้อมูลที่เก็บข้อความที่มีการกำกับข้อมูลอื่นไว้ด้วย (annotated text) คือมีการใส่แท็กต่างๆ กำกับส่วนต่างๆ ในข้อความนั้น เช่น ข้อมูลหมวดคำ ขอบเขตประโยค ตัวอย่างของข้อความที่มีการกำกับด้วยแท็กต่างๆเป็นดังนี้

```
<anthology>
    <poem><title>The SICK ROSE</title>
       <stanza>
            <line>O Rose thou art sick.</line>
            <line>The invisible worm,</line>
            <line>That flies in the night</line>
            <line>In the howling storm:</line>
       </stanza>
       <stanza>
            <line>Has found out thy bed</line>
            <line>Of crimson joy:</line>
            <line>And his dark secret love</line>
            <line>Does thy life destroy.</line>
       </stanza>
    </poem>
</anthology>
```

รูป 1.1 ตัวอย่างข้อมูลที่กำกับตามมาตรฐาน XML

หากพิจารณาจำแนกตามจำนวนภาษา ก็สามารถแบ่งได้เป็นคลังข้อมูลภาษาเดียว (monolingual corpora) หรือคลังข้อมูลหลายภาษา (multilingual corpora) คลังข้อมูลหลายภาษาอาจหมายถึงคลังข้อมูลภาษาที่มีโครงสร้างแบบเดียวกัน เลือกเก็บตัวบทต่างๆเหมือนกัน แต่ทำในคนละภาษา เพื่อให้สามารถเปรียบเทียบความเหมือนความต่างระหว่างภาษาได้ คลังข้อมูลภาษาแบบนี้จะเรียกว่าเป็นคลังข้อมูลเทียบภาษา (comparable corpora) ส่วนคลังข้อมูลหลายภาษาที่เก็บภาษาที่เทียบระหว่างตัวบทภาษาต้นฉบับกับภาษาแปลนั้นจะเรียกว่าคลังข้อมูลเทียบบท (parallel corpora) คลังข้อมูลประเภทหลังนี้จะเป็นประโยชน์ต่อการทำงานของนักแปลและต่อศาสตร์การแปล (translation study) คลังข้อมูล

ประเภทนี้จะต้องมีการจับคู่ (align) ระหว่างข้อความต้นฉบับและข้อความที่แปล เพื่อให้สามารถดึงข้อมูลทั้งสองส่วนออกมาแสดงพร้อมกันได้

6. ความหมายของ Corpus และ Archive

คลังข้อมูลภาษาอาจมีการอ้างถึงได้ในสองลักษณะ กล่าวคือ ในลักษณะแรกจะเป็นการรวบรวมโดยมีการกำหนดไว้ว่าเพื่อใช้เป็นตัวแทนของภาษาประเภทใด ด้วยวัตถุประสงค์อย่างไร การจัดทำคลังข้อมูลภาษาในลักษณะนี้จะเรียกว่า “corpus” ส่วนในลักษณะที่สอง เป็นการจัดทำคลังข้อมูลในลักษณะที่เรียกว่า “archive” ซึ่งเป็นการรวบรวมข้อมูลภาษาในรูปของเอกสารอิเล็กทรอนิกส์ โดยไม่มีการกำหนดเกณฑ์หรือข้อจำกัดใดๆ เพื่อคัดเลือกข้อมูล คลังข้อมูลในลักษณะนี้จึงไม่อาจจัดเป็นตัวแทนของภาษาประเภทใดประเภทหนึ่งโดยเฉพาะ ซึ่งจะพบได้ในคลังข้อมูลภาษาบางแห่งที่เป็นการรวบรวมข้อมูลภาษาตามแต่ที่ผู้จัดทำจะสามารถเข้าถึงแหล่งข้อมูลนั้นได้ (opportunistic) เช่น คลังข้อมูลภาษา Oxford Text Archive ซึ่งจะต่างจากคลังข้อมูลภาษาในลักษณะแรก เช่น คลังข้อมูล Brown หรือคลังข้อมูล SEU ซึ่งเป็นคลังข้อมูลภาษาแบบสมดุลที่ได้กล่าวถึงลักษณะโดยละเอียดไปแล้วในข้างต้น อย่างไรก็ตาม คำว่า corpus นั้น ในความเป็นจริงก็มีการใช้ในลักษณะของความหมายที่สองด้วยคือ หมายถึงข้อมูลภาษาที่รวบรวมขึ้นมาโดยไม่จำเป็นต้องตั้งเกณฑ์แน่นอนชัดเจนอย่างเป็นระบบ ในที่นี้จึงจะใช้คำไทยว่า”คลังข้อมูลภาษา”ในความหมายทั้งสองแบบนี้ และอาจใช้คำที่สั้นลง ”คลังข้อมูล” ในบริบทที่มีการพูดถึงมาก่อนแล้ว

7. คำนวณภาพขนาดของคลังข้อมูล

หากเราต้องการรู้ว่าคลังข้อมูลภาษาขนาดหนึ่งล้านคำ สิบล้านคำ หรือร้อยล้านคำนั้นจริงๆแล้วเป็นข้อมูลภาษาที่มากน้อยเพียงใด เราก็สามารถทำได้

ง่ายๆโดยการสุ่มตัวอย่างหน้าหนังสือเล่มหนึ่ง แล้วนับจำนวนคำที่ปรากฏในหน้าหนังสือนั้น ตัวอย่างเช่น หนังสือ Course in General Linguistics ซึ่งมีจำนวนหน้า 272 หน้า สุ่มไปที่หน้า 81 ของหนังสือเล่มนี้มีจำนวนคำอยู่ 390 คำ หากเราถือว่าทุกๆหน้าในหนังสือเล่มนี้มีจำนวนคำที่ใกล้เคียงกัน ดังนั้น หนังสือเล่มนี้ก็จะมีจำนวนคำประมาณ 106,080 คำ ดังนั้น หากต้องการวัดขนาดคลังข้อมูลภาษาซึ่งมีขนาดหนึ่งล้านคำโดยการเปรียบเทียบกับหนังสือเล่มนี้ ก็จะสามารถคำนวณได้ว่ามีขนาดเท่ากับหนังสือเล่มนี้โดยประมาณ 9.4 เล่ม ดังนั้นจะเห็นได้ว่าคลังข้อมูลภาษาขนาดหนึ่งล้านคำนั้นคิดเป็นสัดส่วนเพียงเล็กน้อยเท่านั้นเมื่อเปรียบเทียบกับจำนวนข้อมูลภาษาที่คนเราพูดหรือเขียนกันในแต่ละวัน ดังนั้น ในการศึกษาที่มีการอ้างถึงการใช้ข้อมูลภาษาเป็นพื้นฐานในการศึกษา การระบุถึงขนาดของคลังข้อมูลที่ใช้จึงเป็นสิ่งสำคัญ เพราะจะทำให้ผู้อ่านสามารถพิจารณาและประเมินความน่าเชื่อถือของผลการศึกษาในงานชิ้นนั้นๆได้ ตัวอย่างเช่น งานวิจัยทางภาษาศาสตร์ที่อ้างว่าศึกษาโดยอาศัยคลังข้อมูลภาษาขนาด 3,000 คำ สามารถตีความได้ว่าใช้ข้อมูลในการศึกษาเท่ากับประมาณ 11 หน้าของหนังสือ ซึ่งข้อมูลเพียงเท่านี้ย่อมไม่สามารถสร้างความน่าเชื่อถือให้กับงานวิจัยนั้นได้มากเท่าที่ควร

8. ปัจจุบันและอนาคตของภาษาศาสตร์คลังข้อมูล

การใช้คลังข้อมูลภาษาจัดว่ามีความสำคัญเป็นอันมากต่อการศึกษาเกี่ยวกับภาษาในปัจจุบัน ทั้งในงานวิจัยทางภาษาศาสตร์ การเรียนการสอนภาษา การแปลภาษา การทำพจนานุกรม และงานวิจัยในด้านการประมวลผลภาษาธรรมชาติ แม้ว่าคลังข้อมูลภาษาในปัจจุบันจะได้เพิ่มจำนวนมากขึ้น รวมทั้งมีขนาดใหญ่กว่าในอดีต อย่างไรก็ตาม ซอฟท์แวร์ (software tool) ที่ใช้ก็ยังคงจำกัดอยู่ที่โปรแกรมคอนคอร์แดนซ์ (concordance) เป็นหลัก ส่วนโปรแกรมอื่นๆ

เช่น โปรแกรมช่วยวิเคราะห์คำอย่าง Word Sketches โปรแกรมช่วยรวบรวมตัวบทสร้างคลังข้อมูลอย่าง CorpusBuilder ฯลฯ ยังไม่ได้รับการเผยแพร่มากนัก หากพิจารณาจากสถานภาพปัจจุบันของการใช้ประโยชน์จากคลังข้อมูลภาษา ก็อาจกล่าวได้ว่ายังมีความต้องการซอฟแวร์ต่างๆ ที่จำเป็นสำหรับการใช้งานร่วมกับคลังข้อมูลภาษาแบบที่มีการกำกับข้อมูล รวมทั้งยังต้องการซอฟแวร์สำหรับช่วยในการกำกับแท็กที่ดี สำหรับในส่วนของการกำกับแท็กในข้อมูลภาษา ซึ่งแต่เดิมยังไม่มีมาตรฐานเดียวกัน และผู้ใช้จำเป็นต้องเข้าใจการกำกับข้อมูลแบบต่างๆที่มีอยู่ในคลังข้อมูลนั้น จึงจะสามารถใช้ประโยชน์จากคลังข้อมูลที่มีการกำกับนั้นได้ ในปัจจุบัน เมื่อมีแนวโน้มว่ามาตรฐานของ SGML และ XML จะเป็นที่ยอมรับสำหรับใช้ในการกำกับข้อมูล ปัญหาต่างๆ ในส่วนนี้จึงลดลง และมีการจัดสร้างคลังข้อมูลภาษา ที่มีการกำกับข้อมูลตามมาตรฐาน SGML และ XML เพิ่มมากขึ้น

บทที่ 2 คลังข้อมูลกับการเรียนการสอนภาษา

1. บทนำ

บทบาทของการใช้คลังข้อมูลภาษาเพื่อการเรียนการสอนนั้นมีมานานแล้ว งานทางด้านคลังข้อมูลภาษาในระยะแรกมีผลอย่างมากต่อการจัดทำหลักสูตรการสอนภาษา (pedagogy) ซึ่งมุ่งให้ความสำคัญกับเรื่องของคำและความถี่ของการเกิดคำนั้น ในช่วงทศวรรษ 1920 ไปจนถึงกลางปีค.ศ. 1950 งานทางคลังข้อมูลภาษาที่เกี่ยวข้องกับการศึกษาคำศัพท์ (vocabulary) และการสอนภาษาในฐานะภาษาที่สองหรือภาษาต่างประเทศของผู้เรียน (second/foreign language pedagogy) จะมีปรากฏค่อนข้างมาก เช่น งานของ Throndike (1921) ที่ใช้ คลังข้อมูลภาษาขนาด 4.5 ล้านคำเพื่อศึกษาความถี่ในการปรากฏของคำในภาษาอังกฤษเพื่อนำมาใช้เตรียมการสอน ซึ่งงานในแนวนี้ได้นำไปสู่การให้ความสำคัญกับประเด็นเรื่องการคัดเลือกคำศัพท์ที่เหมาะสมสำหรับผู้เรียน (vocabulary control) ในการสร้างแบบเรียนการอ่านเป็นเวลาถึง 25 ปี

จากรายงานชื่อ The Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language หรือที่รู้จักในชื่อของ Carnegie Report ในปีค.ศ. 1936 (ซึ่งเป็นผลจากการประชุมทางวิชาการด้านการสอนภาษาอังกฤษในปี ค.ศ.1934 และ 1935 โดยรวมนักภาษาศาสตร์และอาจารย์ผู้สอนภาษาอังกฤษจากที่ต่างๆทั่วโลก) ได้นำเสนอประเด็นสำคัญว่า ในการเรียนการสอนภาษาควรให้ความสำคัญกับเรื่องของความถี่ในการปรากฏของคำ โดยควรเน้นให้ความรู้เกี่ยวกับคำที่มีความถี่ในการปรากฏใช้สูงเป็นหลัก โดย

ต้องพยายามทำให้ผู้เรียนสามารถจดจำติดตัวไปให้ได้ตลอด ให้เป็นความรู้ที่ถาวร (permanent knowledge)

ต่อมาในปี ค.ศ.1944 Thorndike และ Lorge ได้ขยายงานที่ทำไว้ในปี ค.ศ.1921 โดยเพิ่มขนาดของคลังข้อมูลเป็น 18 ล้านคำ นอกจาก Thorndike และ Lorge แล้ว บุคคลสำคัญอีกผู้หนึ่งในยุคนี้ ได้แก่ Michael West (1953) ผู้ซึ่งได้จัดทำ General Service List of English Words ซึ่งถือว่าเป็นฉบับปรับปรุงของ Carnegie Report ในปี ค.ศ.1936 และ West ยังได้ให้ความสนใจกับเรื่องของการนับความแตกต่างทางความหมายของคำแต่ละคำด้วย โดยได้แจกแจงความถี่ของความหมายต่างๆของคำจำนวน 2,000 คำ ซึ่งคัดมาจากรายการคำที่มีความถี่สูงสุด 2,000 คำแรก เช่น ได้พบว่า *mind* มีการใช้ในฐานะคำนามในอัตราส่วนร้อยละ 89 และใช้เป็นคำกริยาร้อยละ 10 ซึ่งก็อาจมีผู้ที่ไม่เห็นด้วยกับผลที่ได้ของคำว่า *mind* นี้ แต่ในที่นี้ สามารถอธิบายได้ว่าเป็นเพราะคลังข้อมูลภาษาที่ใช้นั้นเป็นคลังข้อมูลภาษาเขียนเป็นหลัก

จากช่วงปลายของทศวรรษที่ 1940 เป็นต้นมา ความสนใจในทฤษฎีการสอนภาษาได้เปลี่ยนจากการเน้นความรู้เรื่องคำศัพท์ (vocabulary) มาเน้นที่โครงสร้างภาษา (structural syllabus) ทั้งนี้เป็นผลสืบเนื่องจากแนวคิดของไวยากรณ์โครงสร้างซึ่งพัฒนาขึ้นในช่วงนั้น แนวคิดนี้พิจารณาว่า ภาษามีความสำคัญที่โครงสร้างไวยากรณ์ หากผู้เรียนมีความเข้าใจในโครงสร้างประโยคแบบต่างๆได้ดี ก็จะสามารถเลือกคำที่เหมาะสมในหมวดคำนั้นมาเติมในตำแหน่ง (slot) ที่ว่างของโครงสร้างได้ ต่อมาในช่วงทศวรรษที่ 1960 เมื่อการศึกษาเกี่ยวกับภาษาได้รับอิทธิพลจากแนวคิดเรื่องไวยากรณ์ปริวรรต (Generative grammar) ของ Chomsky เป็นอย่างมาก จึงมีผลให้แนวคิดหลักเกี่ยวกับการสอนภาษา (pedagogic paradigm) เริ่มเปลี่ยนแปลงไปเช่นกัน กล่าวคือ

Chomsky จะให้ความสำคัญกับข้อมูลภาษาจริงน้อยมาก แต่ให้ยึดถือ intution ของเจ้าของภาษาเป็นหลักในการพิจารณาประเด็นความถูกผิดทางไวยากรณ์ ทำให้ประเด็นทางสถิติของการใช้คำต่างๆเริ่มไม่ได้รับความสนใจ และเริ่มมีการพูดถึงข้อจำกัดของการใช้คลังข้อมูลภาษา เช่น คลังข้อมูลส่วนมากเป็นภาษาเขียนมากกว่า คำศัพท์ที่ได้จึงสะท้อนว่าคำอะไรที่ใช้อ่านเขียนมากกว่าที่จะบอกว่าคำอะไรที่เรามักใช้พูดหรือได้ยิน และยังพบว่าธรรมชาติของคลังข้อมูลภาษามีลักษณะการกระจายตัวที่ไม่เสมอกันสูงมาก ตัวอย่างเช่น ในคลังข้อมูลหนึ่ง เราอาจพบว่า คำศัพท์ 4,000-5,000 คำอาจจะกินเนื้อที่ประมาณ 95% ของคลังข้อมูล แต่เมื่อดูคำศัพท์1,000 คำแรกที่ ปรากฏบ่อยที่สุดจะกินเนื้อที่ 85% ของคลังข้อมูล และเมื่อดูคำศัพท์ 50 คำแรกพบว่ากินเนื้อที่ 60% กล่าวคือจะมีกลุ่มคำจำนวนน้อยที่มีการใช้บ่อยมากและครอบครองเนื้อที่ของคลังข้อมูลไปมากกว่ากึ่งหนึ่ง ในขณะที่คำส่วนมากจะปรากฏน้อยหรือปรากฏเพียงครั้งเดียว นอกจากนี้ ยังมีประเด็นว่าคลังข้อมูลภาษาในยุคนี้ยังไม่ได้เป็นตัวแทน (representative) ของภาษาอย่างแท้จริง ตัวอย่างเช่น คลังข้อมูลภาษาของ Thorndike (1921) นั้นประมาณสามในสี่ของข้อมูลเป็นข้อความจากคัมภีร์ไบเบิลและงานวรรณกรรมในศตวรรษที่ 19 จึงไม่ใช่ตัวแทนที่ดีของภาษาอังกฤษทั่วๆไป

นอกจากนี้ แนวคิดในเรื่อง functional-notional เข้ามามีบทบาทมากขึ้นในการเรียนการสอนภาษาที่สอง ทำให้เกิดหลักสูตรการสอนภาษาแบบ functional-notional syllabus ซึ่งเป็นการสอนที่เน้นให้ความสำคัญกับเรื่องของสถานการณ์ (situation) มากขึ้น เกิดเป็นแนวทางใหม่ในการสอนภาษา คือ แนวทางแบบการสอนภาษาเพื่อการสื่อสาร (communicative approach) ซึ่งให้ความสำคัญกับสิ่งที่เป็น communicative competence (Hymes 1981) คือมองว่าความรู้ทางภาษานั้นนอกจากจะมีความรู้ในเรื่องของกฎไวยากรณ์ต่างๆแล้ว ยัง

มีเรื่องรู้ในเรื่องของการใช้ภาษาในสถานการณ์ต่างๆด้วย แนวทางแบบการสอนภาษาเพื่อการสื่อสาร จึงให้ความใส่ใจกับเรื่องอื่นๆที่ไม่เกี่ยวข้องกับภาษาด้วย อย่างเรื่องเทคนิคการสอน เรื่ององค์ประกอบต่างๆที่เกี่ยวข้องกับการเรียน เช่น พื้นความรู้เดิม ทัศนคติ จุดมุ่งหมายของการเรียน เป็นต้น แนวคิดแบบภาษาเพื่อการสื่อสารนี้เป็นที่ยอมรับในเรื่องการศึกษาการเรียนรู้ภาษาที่สอง (second language acquisition) มากว่าสามทศวรรษ เหตุผลหนึ่งที่แนวคิดได้รับการยอมรับอย่างแพร่หลาย เป็นเพราะว่าการเรียนภาษาที่สองนั้น มีเป้าหมายเพื่อให้ผู้เรียนสามารถใช้ภาษาเพื่อพูดคุยติดต่อกันมากกว่าที่จะใช้เพื่ออ่านหนังสือต่างประเทศหาความรู้ประสบการณ์ต่างๆ (โดยเฉพาะอย่างยิ่งในประเทศสหรัฐอเมริกาที่มีความจำเป็นในการสอนภาษาอังกฤษแก่ผู้อพยพ) ประกอบกับมีแนวคิดที่ว่าภาษามีไว้เพื่อใช้สื่อสารทั้งความหมายที่เป็นความหมายในประโยค (propositional meaning) และความหมายที่อิงกับบริบทสังคม (social meaning) ภาษาจึงเป็นเรื่องที่สัมพันธ์กับบริบทการใช้ (situation) และความหมายเป็นสิ่งที่ได้จากการประเมินตีความแลกเปลี่ยนกัน (negotiate) ในสถานการณ์นั้นๆ ไม่ใช่มาจากการผนวกความหมายย่อยๆของคำแต่ละคำ (composition) ผลที่เกิดขึ้นก็คือ อาจารย์ผู้สอนภาษาได้เปลี่ยนบทบาทจากการเป็นศูนย์กลางของความรู้ (source of knowledge) มาเป็นผู้สร้างโอกาสหรือจัดสถานการณ์สำหรับการเรียน (organizer of tasks or opportunities for learning) (Kennedy 1998:281) และมีการให้ความสำคัญกับปัจจัยต่างๆที่จะเอื้อต่อการเรียนรู้ภาษามากขึ้น เช่น พื้นฐานของผู้เรียน ทัศนคติของผู้เรียน เป็นต้น แต่ในอีกมุมหนึ่ง ก็มีผู้มองว่า ได้มีการให้ความสำคัญกับปัจจัยอื่นๆนอกเหนือจากตัวภาษามากจนเกินไป ดังเช่นที่ Sinclair (1998) ได้กล่าวว่า พวกอาจารย์ที่สอนตามแนวการสอนภาษาเพื่อการสื่อสาร (communicative language teaching)

นั้นไม่ได้ให้ความสนใจต่อการอธิบายภาษาที่เกิดขึ้นในปัจจุบัน กลับไปเน้นในเรื่องเทคนิคต่างๆในการเรียนการสอน โดยทำเสมือนว่าข้อเท็จจริงหรือความรู้ความเข้าใจในภาษาอังกฤษนั้นได้ถูกค้นพบสมบูรณ์แล้ว หรือกล่าวอีกนัยหนึ่งก็คือ มัวไปสนใจกับปัจจัยอื่นๆที่ไม่ใช่เรื่องของภาษามากเกินไป จนลืมเรื่องของเนื้อหาภาษาที่ควรจะศึกษา

ในปัจจุบัน คลังข้อมูลภาษาเริ่มกลับมาเป็นที่สนใจอีกครั้งสำหรับการเรียนการสอนภาษาโดยเฉพาะอย่างยิ่งสำหรับการเรียนการสอนภาษาที่สอง อาจารย์ผู้สอนภาษาจำนวนหนึ่งเริ่มกลับมาให้ความสนใจกับสิ่งที่เป็นเนื้อหาภาษาที่เรียน (Kennedy 1998:281) ประกอบกับสาขาทางภาษาศาสตร์คลังข้อมูลมีการพัฒนามากขึ้น คลังข้อมูลภาษาจึงเป็นทรัพยากรที่ดีสามารถให้ข้อมูลภาษาที่เป็นจริงเพื่อนำมาใช้อธิบายลักษณะต่างๆทางภาษา และเป็นหลักฐานชี้ให้เห็นว่าองค์ประกอบ (item) และกระบวนการ (process) ทางภาษาแบบใดที่นักเรียนจะต้องได้พบเห็นบ่อยกว่า ซึ่งองค์ประกอบหรือกระบวนการนั้นก็ควรจะถูกจัดให้เป็นสิ่งที่มีความสำคัญมากกว่าในการสอนและวางหลักสูตร นอกจากนี้ การใช้คลังข้อมูลภาษายังช่วยให้การศึกษาเรื่องของคำปรากฏร่วม (collocation) เป็นไปได้สะดวกมากขึ้น กล่าวโดยสรุป ข้อดีของการใช้คลังข้อมูลภาษา ก็คือ ผู้เรียนสามารถเข้าถึงข้อมูลภาษาได้เป็นจำนวนมาก ทำให้ได้มีโอกาสศึกษาจากฐานที่มาของความรู้ที่กว้างกว่าการศึกษาจากตำราเพียงอย่างเดียว นอกจากนี้ ยังสามารถหาความถี่ในการปรากฏของรูปภาษาที่ศึกษาได้ ซึ่งเป็นสิ่งสำคัญในการเรียนรู้ภาษาที่สอง และเป็นสิ่งที่ไม่สามารถหาได้จากการศึกษาแบบที่ใช้การหยั่งรู้ (intuition) เป็นหลัก[1] การใช้ความถี่ในการอธิบายรูปภาษาในการ

[1] ความจริง ก็มีผู้แย้งและตั้งข้อสงสัยในเรื่องประโยชน์ของการใช้ค่าความถี่ในการ

สอนภาษานั้นเป็นที่ยอมรับกันโดยทั่วไปว่ามีความสำคัญต่อการเรียนรู้ของผู้เรียน เนื่องจากในการเรียนภาษานั้น สิ่งที่ควรเรียนก็คือรูปภาษาที่มีความถี่ในการใช้บ่อย และผู้เรียนเองก็ควรได้มีโอกาสศึกษาภาษาจากข้อมูลที่มาจากการใช้จริง (authentic texts) ในสถานการณ์ต่างๆ ซึ่งก็คือ ต้องอาศัยการใช้ข้อมูลหลักจากคลังข้อมูลภาษานั่นเอง

Murison-Bowie (1996) ได้กล่าวถึงประโยชน์ของการใช้คลังข้อมูลภาษาต่อการทำความเข้าใจเนื้อหาภาษาสำหรับอาจารย์ผู้สอนว่า ช่วยให้เห็นคำที่มักปรากฏร่วมกัน (collocation) เห็นการใช้สำนวนต่างๆ อันจะเป็นประโยชน์ต่อความเข้าใจในเนื้อหาภาษาทั้งในเรื่องของคำ โครงสร้างประโยค และวลีที่ทำหน้าที่แบบคำหรือวลีเสมือนคำ (lexical phrase) ในระดับคำนั้น คลังข้อมูลภาษาช่วยให้เข้าใจความหมายและการใช้ในลักษณะต่างๆของคำได้ดีมากขึ้น ดังจะเห็นได้จากการที่สำนักพิมพ์ทั้งหลายต่างก็จัดทำพจนานุกรมโดยใช้คลังข้อมูลภาษาเป็นพื้นฐาน เช่นที่สำนักพิมพ์ Oxford University Press, Longman and Chambers และมหาวิทยาลัย Oxford มหาวิทยาลัย Lancaster, และ British Library ให้การสนับสนุนการจัดทำคลังข้อมูล BNC ส่วนสำนักพิมพ์ HarperCollins and มหาวิทยาลัย Birmingham ให้การสบับสนุนการจัดทำคลังข้อมูล COBUILD ทั้งนี้เพราะการเขียนคำอธิบายศัพท์โดยใช้ intuition ของ

จัดหลักสูตรการสอนภาษา เช่น Widdowson (1990) บอกว่าคำที่มีหลายความหมายนั้น ความหมายที่ปรากฏบ่อยนั้นมักจะเป็นความหมายเชิงอุปลักษณ์ (metaphor) มากกว่าความหมายแก่น (core) และว่าการเรียนรู้คำโดยเข้าใจความหมายแก่นจะช่วยให้ผู้เรียนอนุมาน (deduce) ความหมายเชิงอุปลักษณ์ได้แต่ผู้เรียนไม่สามารถทำในทิศทางที่กลับกันได้เสมอไป

ผู้เขียนนิยามศัพท์อย่างเดียวนั้นไม่ครอบคลุมเพียงพอ เพราะเป็นการเขียนจากความหมายที่ผู้เขียนสามารถนึกถึงได้เท่านั้น Sinclair (1991) ยังกล่าวถึงข้อเสียของพจนานุกรมที่จัดทำโดยวิธีดั้งเดิมไว้ เช่น มักจะเก็บคำที่ตายไปแล้ว ความหมายใหม่ที่เกิดขึ้นจริงมักจะถูกละเลย ตัวอย่างประโยคก็มักเป็นประโยคที่แต่งขึ้นเอง เป็นต้น ส่วนในระดับโครงสร้างนั้น เนื่องจากการที่คำอาจมีหลายความหมาย ปกติแล้วมักจะมีผลทำให้มีความแตกต่างทางด้านโครงสร้างประโยคด้วย การใช้คลังข้อมูลภาษาจึงช่วยให้เราเห็นความแตกต่างทางโครงสร้างได้ชัดเจนขึ้น ซึ่งอันที่จริงการศึกษาคำโดยไม่สนใจบริบทประโยคที่เกิดนั้นแทบจะเป็นไปไม่ได้ และการศึกษาจากคลังข้อมูลนี่เองที่ช่วยให้เราเห็นว่าสิ่งที่เข้าใจและสรุปกันเป็นกฎไวยากรณ์นั้นบางครั้งก็ไม่มีปรากฏใช้จริง หรือสิ่งที่พบจริง ก็อาจจะไม่ได้มีการสรุปเป็นกฎไวยากรณ์เอาไว้ เช่น Kettermann (1995) ชี้ให้เห็นว่า สิ่งที่ตำราไวยากรณ์เคยกล่าวไว้ว่า ในโครงสร้างแบบ reported speech tense ของส่วนที่เป็น reported action นั้นขึ้นอยู่กับ tense ของกริยาที่เป็น reporting verb นั้นไม่เป็นจริงเสมอไป และในกรณีที่พบว่าไม่เป็นไปตามกฎที่ว่านี้ ความจริงแล้วสามารถอธิบายได้ด้วยกฎอื่น

ส่วนในระดับวลีเสมือนคำ (lexical phrase) นั้น การใช้คลังข้อมูลภาษาช่วยให้เราเห็นคำที่มักปรากฏร่วมกันได้ชัดเจน โดยที่คำมากกว่าหนึ่งคำสามารถนำมารวมกันแล้วใช้เสมือนเป็นหน่วยเดียวกัน เช่น กลุ่มคำที่เป็นสำนวน คำพังเพย นอกจากนี้ การศึกษาเรื่องคำปรากฏร่วม (collocation) ซึ่งเป็นคำที่ดูเหมือนจะมีความสัมพันธ์ต่อกันอย่างใกล้ชิด ความรู้ในเรื่องคำปรากฏร่วมนี้เป็นสิ่งที่จำเป็นสำหรับการสอนภาษาอย่างยิ่ง ทั้งนี้เพราะผู้เรียนภาษามักประสบปัญหาไม่สามารถเลือกใช้คำที่เหมาะสมได้ ทำให้ได้ภาษาที่ใช้ฟังดูไม่เป็นธรรมชาติ ในงานทางด้านภาษาศาสตร์คลังข้อมูล ได้มีผู้ที่สนใจศึกษาเรื่องคำ

ปรากฏร่วมโดยใช้วิธีการทางสถิติ เช่น Church et al. (1991) เสนอให้คำนวณค่า mutual information กับ t-score เพื่อวัดระดับความสัมพันธ์ระหว่างคำ เพื่อตรวจสอบดูว่าคำคู่ใดบ้างที่น่าสงสัยว่าจะเป็นคำปรากฏร่วม

Murison-Bowie (1996) กล่าวสรุปว่า กลุ่มผู้ที่เห็นว่าคลังข้อมูลภาษาสามารถนำมาใช้เป็นประโยชน์ต่อการเรียนการสอนได้ สามารถแบ่งออกได้เป็นสองกลุ่ม คือ พวกที่มองว่าการนำคลังข้อมูลภาษามาใช้เป็นสิ่งจำเป็นและจะเป็นแนวทางใหม่ของการเรียนการสอนภาษา และพวกที่มองว่าการนำคลังข้อมูลภาษามาใช้มีประโยชน์ในลักษณะที่จะมาเสริมการเรียนการสอนในปัจจุบัน

2. การนำคลังข้อมูลภาษามาใช้เป็นหลักในการสอนภาษา

กลุ่มพวกที่เชื่อว่าการใช้คลังข้อมูลภาษาจะเป็นสิ่งที่จำเป็นสำหรับการเรียนการสอนภาษานั้นมองว่าการนำคลังข้อมูลมาใช้จะก่อให้เกิดการเปลี่ยนแปลงวิธีการเรียนการสอนไปเป็นแบบการเรียนรู้จากข้อมูล (data-driven learning) (Johns, 1986, 1988, 1991) และทั้งผู้สอนและผู้เรียนจะต้องเปลี่ยนแปลงบทบาทตัวเองให้เหมาะสมกับวิธีการใหม่นี้

2.1 วิธีการเรียนการสอนด้วยการใช้คลังข้อมูลภาษา

การเรียนด้วยวิธีนี้เป็นการเรียนในลักษณะที่เป็น “Data-driven learning” คือเป็นการเรียนภาษาจากคลังข้อมูลภาษาที่มี ดังนั้น ในการสร้างคลังข้อมูลภาษาสำหรับการเรียนการสอน จึงจำเป็นต้องมีการคัดเลือกตัวข้อมูลภาษาที่จะนำมาใช้ รวมทั้งยังต้องเตรียมข้อมูลให้มีปริมาณที่มากพอสำหรับนำมาอนุมานกฎเกณฑ์ของการใช้ภาษาได้ อย่างไรก็ตาม มีข้อควรคำนึงสำหรับผู้ใช้หรือผู้เรียนประการหนึ่ง คือ เนื่องจากข้อมูลภาษาที่มีในคลังข้อมูลนั้นจัดว่าเป็นเพียงส่วนหนึ่งของการใช้ภาษา เป็นเพียงตัวอย่างบางส่วนที่ถูกนำมาแสดงให้เห็น

เท่านั้น ดังนั้น จึงควรระมัดระวังในการใช้ข้อมูลนั้นตีความและสรุปว่าเป็นการใช้ในทุกบริบทของภาษา

ข้อดีของการเรียนด้วยวิธีนี้ก็คือ ผู้เรียนสามารถเข้าถึงข้อมูลได้โดยตรง มิใช่เป็นการเรียนจากสิ่งที่มีผู้อื่นสรุปมาให้ในตำราดังเช่นที่เป็นอยู่ในการเรียนการสอนปัจจุบัน ซึ่งมักจะมีข้อเสียก็คือว่าผู้เขียนตำรามักจะเขียนขึ้นจากกฎเกณฑ์ต่างๆ เท่าที่จะสามารถนึกคิดเอาได้ และถ่ายทอดสืบกันไป เมื่อเวลาผ่านไปนานๆก็อาจจะไม่ได้คัดเลือกเอารูปภาษาที่ไม่ใช้แล้วออกไป ทำให้เกิดข้อจำกัดจากการเรียนรู้จากตำรา คือผู้เรียนจะไม่สามารถแยกแยะรูปภาษาที่เป็นรูปแบบที่ใช้ปกติในปัจจุบัน (common use) กับรูปภาษาที่เลิกใช้กันแล้ว (out of use) ได้ การใช้คลังข้อมูลภาษาอาจนำมาช่วยเสริมในกรณีนี้ได้ โดยทำให้ผู้เรียนสามารถสร้างภาพความเข้าใจในแง่การใช้ตามบริบทต่างๆ ได้ด้วยตนเอง

การศึกษาภาษาด้วยการใช้คลังข้อมูล จำเป็นต้องอาศัยการใช้โปรแกรมคอนคอร์แด๊นซ์ช่วยในการเรียกข้อมูลที่ต้องการขึ้นมา โปรแกรมคอนคอร์แด๊นซ์นี้จะช่วยค้นหาข้อมูลภาษาที่ต้องการพร้อมทั้งแสดงบริบทที่เกิดร่วมกับคำๆนั้นออกมาด้วย (ดูบทที่ 11) และผู้ใช้(ผู้เรียน)สามารถจัดเรียง (sort) ผลที่เรียกขึ้นมาได้ด้วยตนเอง ทั้งจากบริบทที่อยู่ข้างหน้าของคำที่ต้องการ (left sort) หรือจากบริบทข้างหลังของคำ (right sort) ซึ่งจะเป็นประโยชน์อย่างมากต่อการศึกษาเรื่องคำ (word) หรือหน่วยคำ (morpheme) ที่มีหลายความหมายหรือมีหลายหน้าที่ กล่าวคือ เมื่อผู้เรียนสามารถเข้าใจความหมายหลักของคำนั้นๆ แล้ว ผู้เรียนก็จะสามารถใช้คลังข้อมูลภาษาในการค้นคว้าความหมายหรือหน้าที่อื่นๆ ของคำนั้นได้ต่อไป หรือในกรณีที่คำหนึ่งๆ มีรูปแปรได้หลายรูป การที่ผู้เรียนสามารถศึกษาจากตัวอย่างจริงของภาษาได้ ผู้เรียนก็จะได้เห็นความแตกต่างของรูปแปรต่างๆ ที่มีได้ชัดเจนมากขึ้น รวมทั้งเห็นวิธีการใช้และความสัมพันธ์ระหว่าง

รูปแปรของคำคำเดียวนั้น ซึ่งจะช่วยให้เข้าใจการใช้คำนั้นได้ถูกต้องชัดเจนยิ่งขึ้น ในลักษณะนี้ ก็อาจกล่าวได้ว่า การใช้คลังข้อมูลภาษา นับเป็นอีกรูปแบบหนึ่งของวิธีการศึกษาที่จะเป็นประโยชน์ได้ต่อไป

2.2 บทบาทของผู้เรียนและผู้สอนในชั้นเรียน

ดังที่ได้กล่าวมาในตอนต้น การใช้คลังข้อมูลภาษาสำหรับการเรียนการสอนภาษานั้น จำเป็นต้องมีโปรแกรมคอนคอร์แด๊นซ์และคลังข้อมูลภาษาที่เหมาะสม ทั้งนี้เพื่อให้ทั้งผู้สอนและผู้เรียนได้มีโอกาสค้นคว้าเกี่ยวกับการใช้ภาษาในลักษณะต่างๆ ได้มากและหลากหลาย ดังนั้น โดยลักษณะของการเรียนการสอนแบบนี้ ผู้เรียนจะมีบทบาทเสมือนเป็นนักวิจัยทางภาษา อาจารย์ผู้สอนเองก็จะเปลี่ยนบทบาทจากผู้ถ่ายทอดความรู้ไปเป็นผู้ประสานงานคอยให้ความช่วยเหลือแก่นักเรียนให้กระทำดังกล่าวให้สำเร็จลุล่วง ซึ่งจะทำได้ก็ต้องอาศัยความพร้อมในหลายๆด้าน ในด้านอาจารย์ผู้สอนเองนอกจากความรู้ความเข้าใจในภาษานั้นๆ แล้ว ก็จะต้องมีความรู้ความเข้าใจเกี่ยวกับการเรียนในแบบใหม่นี้ ต้องรู้ว่าจะใช้โปรแกรมคอนคอร์แด๊นซ์ช่วยค้นหารูปแบบตัวอย่างการใช้ได้อย่างไรบ้าง จะวิเคราะห์หรือหาข้อสรุปลักษณะร่วม (generalization) จากผลลัพธ์ที่ได้จากโปรแกรมคอนคอร์แด๊นซ์อย่างไร ส่วนผู้เรียนก็เช่นกัน จะต้องได้รับการฝึกให้ใช้โปรแกรมคอนคอร์แด๊นซ์ได้ รู้จักหาแหล่งของคลังข้อมูลภาษา และอาจจะต้องฝึกฝนให้สามารถมองเห็นข้อสรุปที่เป็นลักษณะร่วมทางภาษาจากข้อมูลที่ให้ หลังจากเตรียมความพร้อมนี้แล้ว ในทางปฏิบัติ ผู้สอนก็อาจให้ผู้เรียนทำการวิเคราะห์จากคลังข้อมูลภาษาด้วยตัวเองได้ ซึ่งอาจเป็นในรูปของการมอบหมายหัวข้อเฉพาะเรื่องหนึ่งๆ ให้ทำการศึกษา หรืออาจให้ผู้เรียนเลือกหัวข้อที่สนใจศึกษาขึ้นมาเอง ซึ่งการกำหนดให้ผู้เรียนเป็นผู้ใช้คลังข้อมูลในการศึกษาภาษาเอง จะช่วยให้บทบาทของผู้เรียนมีลักษณะที่ริเริ่มเอง (active) และสร้างสรรค์

(creative) มากขึ้น

กล่าวโดยรวมแล้ว จะเห็นได้ว่า การนำคลังข้อมูลภาษาเข้ามาใช้จะทำให้บทบาทของผู้เรียนและผู้สอนเปลี่ยนไปจากเดิม โดยผู้เรียนจะเป็นเสมือนนักวิจัย คือต้องเรียนรู้จากการปฏิบัติและค้นคว้าด้วยตนเองมากขึ้น และต้องประมวลความรู้จากข้อมูลภาษาที่ได้มาเอง โดยบางครั้งก็อาจขอความช่วยเหลือจากผู้สอนบ้าง และในการนี้ บทบาทของผู้สอนเองก็จะลดการเป็นผู้ให้ข้อมูลในลักษณะของการสอนลง แต่จะกลายมาเป็นผู้ให้ความช่วยเหลือและชี้แนะแนวทางแก่ผู้เรียน ซึ่งคาดว่าจะเป็นการดีสำหรับผู้เรียนที่จะได้ฝึกทักษะการค้นคว้าและศึกษาด้วยตนเองมากขึ้น อันจะนำไปสู่การพัฒนาทักษะในการเรียนรู้ในสาขาอื่นๆ ให้ดียิ่งขึ้นไปด้วย

3. การประยุกต์ใช้คลังข้อมูลภาษาในการสอนภาษา

แม้ว่าแนวทางข้างต้นที่ให้นำคลังข้อมูลภาษามาใช้เพื่อการเรียนภาษาโดยตรงนั้นจะเป็นประโยชน์ในหลายด้าน แต่ในทางปฏิบัติอาจเป็นไปได้ยาก เพราะจะต้องมีการเปลี่ยนแปลงทั้งรูปแบบการเรียนสอนและทัศนคติเกี่ยวกับการเรียนการสอนใหม่ จึงมีผู้เสนอว่า แทนที่จะใช้คลังข้อมูลภาษาในการเรียนการสอนภาษาโดยตรงเลย ควรจะนำคลังข้อมูลภาษามาประยุกต์ใช้กับการเรียนการสอนภาษาน่าจะเหมาะสมกว่า การพิจารณาว่าจะนำคลังข้อมูลภาษามาประยุกต์ใช้กับการเรียนการสอนอย่างไรนั้น Widdowson (1980) มองว่าสามารถพิจารณาได้จากมุมมองสองแบบ ขึ้นอยู่กับการตีความคำว่าภาษาศาสตร์ประยุกต์ (applied linguistics) ว่ามีความหมายอย่างไร ในมุมมองแบบแรก หากภาษาศาสตร์ประยุกต์หมายถึงการนำความรู้ทางภาษาศาสตร์มาประยุกต์ใช้ บทบาทของคลังข้อมูลภาษาก็จะเป็นเสมือนแหล่งข้อมูลที่ช่วยให้เราสามารถอธิบายปรากฏการณ์ต่างๆในภาษาได้ดีขึ้น ช่วยให้สามารถเขียนพจนานุกรมหรือ

ตำราไวยากรณ์ได้ดีขึ้นอันจะเป็นประโยชน์ต่อการเรียนการสอนภาษาอีกทอดหนึ่ง ทั้งยังช่วยให้สามารถวางหลักสูตรแนวการสอนได้ดีขึ้นเพราะสามารถเลือกคำศัพท์และโครงสร้างไวยากรณ์ที่เหมาะสมสำหรับการเรียนในลักษณะที่สอดคล้องกับการใช้ภาษาจริงๆได้ ซึ่งสิ่งที่ค้นพบจากงานวิจัยที่ศึกษาโดยอาศัยคลังข้อมูลหลายๆงานเป็นตัวอย่างที่แสดงให้เห็นว่าความเชื่อแต่ดั้งเดิมในบางเรื่องที่เราใช้สอนภาษากันต่อๆมานั้นอาจจะไม่ถูกต้องเสียทีเดียว หรือหากเราศึกษาคลังข้อมูลของเอกสารเฉพาะทางก็จะช่วยให้เราเห็นถึงลักษณะที่แตกต่างกันของตัวบทแต่ละประเภทได้ ดังในตัวอย่างงานวิจัยของ Biber (1988) ที่ได้แสดงความแตกต่างของตัวบทแต่ละประเภทจากการพิจารณาคุณสมบัติทางภาษาในมิติต่าง ๆ ซึ่งผลที่ได้จะแสดงให้เห็นถึงระดับความเหมือนและความต่างของตัวบทแต่ละประเภท ช่วยให้สามารถจัดทำหลักสูตรของ ESP (English for Specific Purpose) ได้ดีมากขึ้น ส่วนที่มีผู้มองว่าสามารถนำเอาโปรแกรมคอนคอร์แด็นซ์และคลังข้อมูลภาษามาใช้เป็นทรัพยากรการเรียนโดยตรงในลักษณะที่เป็นการเรียนรู้จากข้อมูล (data-driven learning เช่น Johns 1986, 1988, 1991) นั้น Sinclair (1991) แย้งว่าการนำวิธีการวิจัยที่นักภาษาศาสตร์คลังข้อมูลใช้มาใช้กับนักเรียนโดยตรงนั้นอาจจะไม่เหมาะสม เนื่องจากผู้เรียนอาจจะไม่สามารถหาคลังข้อมูลขนาดใหญ่ใช้ได้สะดวกและการวิเคราะห์ข้อมูลโดยอาศัยเพียงความรู้ของเจ้าของภาษาอย่างเดียวอาจไม่เพียงพอ เพราะผู้เรียนเหล่านี้อาจจะยังไม่มีความรู้ทางด้านภาษาศาสตร์เพียงพอสำหรับวิเคราะห์ข้อมูล ดังนั้นจึงควรใช้คลังข้อมูลภาษาเป็นแหล่งทรัพยากรให้อาจารย์ผู้สอนสามารถใช้เพื่อเตรียมเนื้อหาและหลักสูตรสำหรับใช้ในการสอนภาษาอีกต่อหนึ่ง

ส่วนในอีกมุมมองหนึ่งของการนำคลังข้อมูลภาษามาประยุกต์ใช้ เป็นการมองภาษาศาสตร์ประยุกต์ในลักษณะของการจัดหลักสูตรสอน (pedagogic

context) เป็นหลัก การที่โปรแกรมคอนคอร์แด๊นซ์และคลังข้อมูลภาษาสามารถช่วยให้เราอธิบายปรากฏการณ์ทางภาษาได้ดีขึ้นไม่ได้เป็นหลักประกันว่าจะสามารถนำผลที่ได้นั้นมาใช้ให้เป็นประโยชน์ต่อการเรียนการสอนได้เสมอไป ทั้งนี้เพราะเราต้องพิจารณาจากกระบวนการ (process) ของการเรียนการสอนเพื่อดูว่าความรู้ที่ได้จากการศึกษาคลังข้อมูลภาษานั้นเกี่ยวข้องหรือเหมาะสมต่อการเรียนการสอนเพียงใด อีกทั้งการวิเคราะห์ภาษาจากรายการคำเรียงที่ได้จากโปรแกรมคอนคอร์แด๊นซ์เพียงอย่างเดียวนั้นเป็นการตัดตอนภาษาออกจากบริบทการสื่อสาร ทำให้ผู้เรียนไม่สามารถเห็นถึงกระบวนการตีความและหาความหมายจากวัจนกรรมที่เกิดขึ้น ซึ่ง Widdowson (1984, 1990) ถือว่าเป็นเรื่องสำคัญของการเรียนรู้ภาษา (language acquisition) และขัดกับหลักการสำคัญของการเรียนภาษาเพื่อการสื่อสาร (communicative approach) ที่เน้นความสำคัญของการเรียนรู้จากการได้ใช้ แต่ทั้งนี้ก็ไม่ได้หมายความว่าคลังข้อมูลภาษาจะไม่เกี่ยวข้องกับการเรียนการสอนภาษา แต่เป็นสิ่งที่ผู้วางหลักสูตรจะต้องตัดสินว่าผลที่ได้จากการศึกษาผ่านคลังข้อมูลแบบไหนที่สามารถนำมาใช้ให้เหมาะสมกับหลักสูตรที่วางไว้ ซึ่ง Aston (1995) เป็นผู้หนึ่งที่มองว่า ผลที่ได้จากการใช้โปรแกรมคอนคอร์แด๊นซ์กับคลังข้อมูลภาษานั้นสามารถนำมาปรับใช้ให้เข้ากันได้กับการเรียนการสอนในแบบภาษาเพื่อการสื่อสาร (communicative approach)

แนวคิดทางด้านภาษาศาสตร์คลังข้อมูลและการจัดทำหลักสูตรสอนภาษา (language pedagogy) มีลักษณะที่เหมือนกันบางประการ คือทั้งสองแนวคิดมองภาษาทั้งในเรื่องของการสร้างรูปประโยค (production) และเรื่องของการตีความ (interpretation) ว่ามีลักษณะที่เป็นการนำคำหรือวลีที่จำไว้เป็นส่วนๆ (ready-made memorized building blocks) มาใช้มากกว่าที่จะเป็นการนำคำแต่ละคำมารวมกัน (composition) ตามกฎเกณฑ์ไวยากรณ์เพียงอย่างเดียว

แนวคิดแบบนี้มองสามัตถิยะภาษาและการใช้ภาษาในลักษณะที่เป็นแบบมีรูปแบบสำเร็จหรือมีแบบแผนการใช้ (idiomatic or schematic view of language competence and use) การใช้กลุ่มคำหรือวลีที่จดจำไว้นี้เป็นแนวความคิดที่แฝงอยู่ในวิธีการสอนแบบภาษาเพื่อการสื่อสาร เพราะการสอนภาษาด้วยวิธีนี้ให้ความสนใจกับสิ่งที่เกิดขึ้นจริงมากกว่าที่จะสนใจความถูกต้องทางไวยากรณ์ (grammatically correct) หรือความสอดคล้องกับการรับรู้ (psychologically feasible) ในทางภาษาศาสตร์คลังข้อมูลก็มองภาษาในลักษณะที่คล้ายกัน คือ มองภาษาว่าไม่ได้เกิดจากการนำแต่ละคำมารวมกันตามเกณฑ์ไวยากรณ์เพียงอย่างเดียว แต่มีในลักษณะที่มีการปรากฏร่วมกันของคำจนเป็นเหมือนรูปแบบสำเร็จ (idiomatic colligation) Sinclair (1991) มองว่าในการตีความ (interpret) ข้อความใดๆ จำเป็นต้องใช้ทั้งหลักการเลือกใช้รูปแบบสำเร็จ (idiom principle) และหลักการเลือกใช้กฎ (open choice principle) โดยที่หลักการเลือกใช้รูปแบบสำเร็จมองว่าคนเรามีรูปวลีกึ่งสำเร็จรูป (semi-preconstructed phrases) ที่พร้อมใช้อยู่เป็นจำนวนมาก ส่วนหลักการเลือกใช้กฎมองว่าภาษาเป็นเรื่องของการเลือกใช้คำหรือวลีหรือประโยคที่เหมาะสมตามข้อกำหนดทางไวยากรณ์ โดยที่ Sinclair มองว่าหลักที่สำคัญที่สุดคือหลักการเลือกใช้รูปแบบสำเร็จเพราะเป็นหลักที่ใช้ในการตีความตามปกติ ยกเว้นในบางกรณีที่จะเปลี่ยนไปใช้หลักการเลือกใช้กฎ นอกจากนี้ ยังมีความเชื่อกันว่า ในวลีหรือประโยคเดียวกันสามารถสร้างขึ้นได้จากการผสมคำตามหลักไวยากรณ์หรือจากการใช้รูปแบบสำเร็จ (schematic) ซึ่งแต่ละคนสามารถปรับเปลี่ยนองค์ความรู้ที่มีอยู่ได้ โดยวลีหรือประโยคที่เกิดจากการรวมแต่ละคำตามกฎเมื่อมีการใช้หรือได้ยินจนเป็นปกติวิสัยก็จะปรับขึ้นเป็นรูปแบบสำเร็จได้ ซึ่งในความเชื่อแบบนี้ การใช้คลังข้อมูลภาษาจะเป็นประโยชน์โดยตรงเพราะช่วยให้ผู้เรียนได้เห็นรูปแบบการ

ใช้ที่เกิดซ้ำๆกันนี้ได้ ทำให้สามารถเรียนรู้รูปแบบสำเร็จ (acquire schematic knowledge) การใช้โปรแกรมคอนคอร์แด็นซ์และคลังข้อมูลจึงช่วยให้เรารู้ว่าอะไรที่มีการปรากฏใช้บ่อย และยังช่วยให้ศึกษาเรื่องของสำนวนตายตัว (fixed phrase) และเรื่องคำปรากฏร่วม (collocation) ได้ง่ายมากขึ้น แต่อย่างไรก็ตาม Aston (1995) เสนอว่า คลังข้อมูลภาษาที่จะนำมาใช้เพื่อประกอบการสอนไม่จำเป็นต้องเป็นคลังข้อมูลขนาดใหญ่ที่เป็นตัวแทนภาษาทั้งหมดเหมือนกับที่นักภาษาศาสตร์คลังข้อมูลต้องการใช้ ทั้งนี้เนื่องจากนักภาษาศาสตร์ต้องการอธิบายสิ่งที่เป็นปรากฏการณ์ทางภาษาในภาพรวม แต่ในเรื่องการสอนภาษาถ้ามองว่าการเรียนรู้ภาษาของคนนั้นเป็นการเรียนจากข้อมูลที่ได้ทีละส่วน รูปแบบสำเร็จ (schemata) ที่ได้จึงมีการปรับเปลี่ยนได้ตลอด คลังข้อมูลที่นำมาใช้เพื่อสอนภาษาจึงมีขนาดเล็กได้ แต่ควรเป็นข้อมูลชนิดเดียวกัน เพื่อให้เห็นรูปแบบการเกิดที่ซ้ำๆกันในบริบทแบบเดียวกัน แต่อย่างไรก็ตาม ในมุมมองของการสอนแบบภาษาเพื่อการสื่อสารนี้ แม้ว่าผู้เรียนจะสามารถมองเห็นลักษณะร่วมในทางภาษา (regularity) จากคลังข้อมูลที่ให้ ก็ไม่ได้เป็นหลักประกันว่าผู้เรียนจะสามารถใช้ความรู้นั้นในการติดต่อสื่อสารได้ ดังนั้น Aston (1995) จึงเสนอให้จัดงาน (task) ของการวิเคราะห์คลังข้อมูลให้เป็นงานหนึ่งที่อยู่ภายในแบบฝึกหัดการสื่อสาร (communicative task) ซึ่งอาจทำได้ในสองลักษณะ วิธีแรกเป็นการใช้คลังข้อมูลภาษาเพื่อเป็นเครื่องมืออ้างอิง (reference tool) สำหรับใช้แก้ปัญหาที่เกิดขึ้นในแบบฝึกหัด (communicative task) นั้นๆ เช่น ใช้คลังข้อมูลเฉพาะสาขาที่รวบรวมไว้เพื่อเป็นตัวอย่างเวลาที่สอนแปลเอกสารในลักษณะเดียวกัน ส่วนวิธีการที่สองคือใช้คลังข้อมูลภาษาเป็นแหล่งข้อมูลสำหรับการทำแบบฝึกหัด ซึ่งคลังข้อมูลภาษาจะเป็นประโยชน์อย่างมากต่อการทำกิจกรรมการอ่าน (reading activity) เช่น นำเอาเอกสารมาใช้ประกอบกิจกรรม

การอ่านแบบตัวต่อ (jigsaw reading) นอกจากนี้ อาจารย์ผู้สอนภาษาอาจใช้ประโยชน์จากคลังข้อมูลภาษาและโปรแกรมคอนคอร์แด๊นซ์เพื่อใช้เตรียมแบบฝึกหัดต่างๆสำหรับการเรียนการสอนภาษาได้ ดังที่ Tribble และ Jones (1990) ได้แสดงให้เห็นถึงกลวิธีต่างๆที่อาจารย์ผู้สอนจะสามารถนำผลที่ได้จากการใช้โปรแกรมคอนคอร์แด๊นซ์มาเป็นประโยชน์ เช่น ใช้ผลที่ได้เพื่อทำแบบฝึกหัดเติมคำในช่องว่าง ใช้เพื่อให้นักเรียนหาความหมายต่างๆของคำที่ได้จากการทำคอนคอร์แด๊นซ์ ใช้เพื่อให้นักเรียนศึกษาเรื่องของคำหลายหน้าที่โดยให้หาหมวดคำของคำนั้นๆในบริบทต่างๆที่กำหนด ใช้เพื่อศึกษาเรื่องของคำที่เป็นคำพ้องความหมายหรือมีความหมายใกล้เคียงกัน เช่น ศึกษาว่าคำว่า *over* และ *above* ว่ามีความเหมือนหรือแตกต่างกันอย่างไรบ้าง เป็นต้น ตัวอย่างในรูปที่ 2.1 เป็นเพียงส่วนหนึ่งของแบบฝึกหัดที่สามารถจัดทำขึ้นได้โดยการใช้โปรแกรมคอนคอร์แด๊นซ์กับคลังข้อมูลภาษาที่มี (ดู Mallikamas 1999)

Exercise 1 (gapfill exercise) Use the context clues to identify the missing keyword in the following list.

```
*t several times, and yet, when I ask you to ..... a trip with me, to act ballast-
*the wireless operator on the grounds would  ..... any message that came to him.
*t that  his craft was working well, let Ned ..... charge and devoted him to seein
*d with a smile, "I think I should prefer to ..... my first ride in your larger sh
*ds could form a notion of what was about to ..... place, Tom had sent his craft t
*have a plan partly thought out, but it will ..... some time to finish it.
*eedy, and leaving Tom and Captain Weston to ..... the first watch, the others wen
*who naturally acted as guide, "as it would  ..... too long to put together and be
```

The missing keyword is .....

Exercise 2 (the study of parts of speech) According to the data given below, how many parts of speech can the word *round* be?

```
1  have turned him off, he coolly turned round and said that my father was a ge
2  outh where the sun shone all the year round and he could ride half the day w
3  lue sky; smooth and dimpled slopes as round as cones, bare but for the green
4  in a brown hat and blue dress, with a round face and snub nose went and bo..
5  t the neighborhood and Jo was looking round for help when Laurie, who had he
6  , I rather like it, for she winds one round her finger as softly and prettil
7   electrified him by throwing her arms round his neck, and crying out, with a
8  ch untied the shoes which he had hung round his neck gave way, and both fell
9  ade with a very full skirt and little round jacket, but embroidered by her o
10        Then they would select a long, round stick --which must be also perfe
11 s had an ingenious way of fastening a round stone to the end of a long handl
12 I can't induce him to give me a good, round sum of account--fifty thousand,
13 pestle another stone was a smooth and round surface at the bottom was used.
14 quickly, he saw Micky Maguire running round the corner of the street which t
15 me spread the table, set the children round the fire and fed them like so ma
```

```
16 h her broom she soon swept a path all round the garden, for Beth to walk in
17 descended from his throne, and coming round to the side opened the door and
18 y met with a troop of women collected round the well, talking and laughing.
19 self tossing on the ocean in a voyage round the world, when the sound of voi
20 l be crying so hard that the tick fog round you will obscure the prospect.
21  to last till the silver wedding came round , and invented three different k
22  we pass," said Jo, as the plate went round , and  the ice began to melt out
23 is year the summer floods brought the round , brown, fruitful cones to my v
```

Exercise 3 (the study of grammatical features) According to the data given, what are the 4 different meanings of *can*?

```
1          The man must be a prodigy who can restrain his manners and morals, undep
2  ing to the pages of the dramatist, we can picture to ourselves the greatness of
3  ng of the twentieth century barbarism can throw off its gentle disguise, and bur
4                                         Can you tell me, sir," he inquired, turnin
5           No American, living or dead, can surpass the verbal infelicity of Knatc
6  ne out on an errand, she said," and I can find no better time for telling you wh
7                   In the meantime, you can invite him an Grace to come and stay w
8  ing to pieces the girl shows what she can do to support the family.
9                                        Can you show me the way to Broadway?
10 fteen dollars a week, you know, and I can get along on ten, but of the five I sa
11          "Perhaps," thought Frank, "he can help me to some employment for my even
12            I will go down and see if I can find something to eat.
14               If I am in the city, and can afford to buy a dress-suit, I might co
15                                    You can go.
16                   No, at any rate, I can wait, since she is to be mine at last.
17                                       Can I have some water to wash my face?
18                        "All right, he can keep them."
19 go with me as far as Broadway, then I can get into an omnibus.
20                 But honestly, I know I can do a good deal better than I am doing
```

บทที่ 3 คลังข้อมูลกับการวิจัยภาษา

1. บทนำ

ในงานการศึกษาวิจัยที่เกี่ยวกับภาษา Biber, Conrad, และ Reppen (1996) มองว่าโดยทั่วไปเป็นการศึกษาวิจัยในสองลักษณะคือ การศึกษาด้านโครงสร้างภาษา (language structure) และการศึกษาด้านการใช้ภาษา (language use) การใช้คลังข้อมูลภาษาจะช่วยทำให้เห็นถึงโครงสร้างและการใช้ภาษาได้ดีขึ้น ซึ่งแต่เดิมมาการอธิบายเกี่ยวกับโครงสร้างภาษาจะเป็นไปในแบบดั้งเดิมที่ไม่ได้ใช้วิธีวิจัยเชิงประจักษ์และยึดติดกับความหยั่งรู้ (intuition) ของผู้วิจัย แต่ในปัจจุบัน นักวิจัยสามารถนำคลังข้อมูลภาษาเข้ามาใช้เพื่อช่วยอธิบายโครงสร้างภาษาได้ ซึ่งการใช้คลังข้อมูลภาษาจะช่วยให้นักวิจัยได้เห็นตัวอย่างภาษาจริงๆ รวมทั้งยังสามารถชี้ให้เห็นถึงโครงสร้างประโยคบางอย่างที่มักไม่อยู่ในความสนใจศึกษาของทั้งผู้เรียนและผู้สอน กล่าวคือผู้วิจัยจะสามารถสังเกตเห็นถึงการปรากฏของรูปโครงสร้างที่ดูคล้ายคลึงกัน (seemingly similar structures) แต่ปรากฏในบริบทและในหน้าที่ที่ต่างกัน อย่างไรก็ตาม การใช้ประโยชน์จากคลังข้อมูลภาษาที่ดูจะมีอิทธิพลมากที่สุด คือการใช้ประโยชน์ในการศึกษาวิจัยเรื่องการใช้ภาษา (language use) ซึ่งเป็นการศึกษาที่เน้นที่การใช้ภาษาที่เกิดขึ้นจริง (actual practice) และปริมาณการใช้ของรูปภาษาหนึ่งๆ (linguistic pattern) ว่ามีการใช้มากหรือน้อยเพียงใด

ในปัจจุบัน มีงานหลายงานที่ให้ความสำคัญและใช้ประโยชน์จากคลังข้อมูลภาษา เช่น งานด้านการทำพจนานุกรมหรือการศึกษาเกี่ยวกับคำศัพท์

การใช้คลังข้อมูลภาษาจะช่วยให้ผู้วิจัยเห็นขอบเขตของความหมายต่างๆ (range of meaning) ของคำได้ครอบคลุมมากขึ้น ทำให้สามารถเข้าถึงความหมายใหม่ๆ ได้โดยมีหลักฐานจากการใช้ที่มีปรากฏจริงอยู่ในตัวบท

งานวิจัยทางด้านการใช้ภาษานั้นจำเป็นต้องเป็นการศึกษาเชิงประจักษ์และอ้างอิงข้อมูลจริง ซึ่งสิ่งเหล่านี้เป็นสิ่งที่ได้จากการใช้คลังข้อมูลภาษา เพราะการศึกษาภาษาโดยอาศัยคลังข้อมูลภาษาในปัจจุบันนั้นจะมีลักษณะที่สำคัญดังต่อไปนี้

- เป็นการศึกษาเชิงประจักษ์จากภาษาที่เกิดขึ้นจริง
- เป็นการศึกษาแบบอิงคอมพิวเตอร์ (computer-based) ทั้งในแบบที่ให้คอมพิวเตอร์ช่วยวิเคราะห์โดยอัตโนมัติหรือแบบที่มีปฏิสัมพันธ์กับผู้ใช้ (automatic and interactive techniques)
- วิธีการศึกษาเป็นได้ทั้งการวิเคราะห์แบบเชิงคุณภาพ (qualitative) และแบบเชิงปริมาณ (quantitative)

การใช้คลังข้อมูลภาษาสามารถช่วยให้นักวิจัยได้ค้นพบเรื่องใหม่ๆ ในการศึกษาภาษาได้ โดยเฉพาะอย่างยิ่งการศึกษาแบบแผนความความสัมพันธ์ (association patterns) (Biber et al. 1996:116) ซึ่งแบบแผนความความสัมพันธ์ดังกล่าว Biber et al. เสนอว่าเราสามารถมองได้ใน 2 แบบ คือ หนึ่งเป็นแบบแผนความความสัมพันธ์นอกภาษา (non-linguistic association patterns) ได้แก่ การศึกษาความสัมพันธ์ของลักษณ์ทางภาษา (linguistic feature) กับสิ่งที่ไม่ใช่ลักษณ์ทางภาษา (non-linguistic feature) เช่น ดูการปรากฏของลักษณ์ทางภาษาบางอย่างตามทำเนียบภาษา (registers) ตามถิ่น (dialects) หรือตามประวัติการเปลี่ยนแปลง (historical change) ส่วนแบบที่สอง เป็นแบบแผนความความสัมพันธ์ภายในภาษา (linguistic association

patterns) ได้แก่การศึกษาลักษณ์หนึ่งในภาษา (linguistic feature) เทียบกับอีกลักษณ์หนึ่งในภาษา ซึ่งก็สามารถแยกได้เป็นแบบแผนความสัมพันธ์ระดับคำ (lexical associations) และแบบแผนความสัมพันธ์ระดับไวยากรณ์ (grammatical associations)

สำหรับการศึกษาลักษณ์ทางภาษา (linguistic feature) เราสามารถศึกษาได้ทั้งในระดับคำและในระดับโครงสร้าง ในกรณีของการศึกษาในระดับคำ หากเป็นการศึกษาแบบแผนความสัมพันธ์ระดับคำ (lexical associations) ก็จะเป็นการศึกษาเรื่องความสัมมพันธ์ระหว่างคำนั้นกับคำอื่นๆ หรือศึกษาการปรากฏร่วมกันของคำ (collocation) และหากเป็นการศึกษาแบบแผนความสัมพันธ์ระดับไวยากรณ์ (grammatical associations) ก็จะเป็นการศึกษาเรื่องความสัมพันธ์ของคำนั้นกับโครงสร้างประโยค คือศึกษาโครงสร้างที่คำนั้นมักจะปรากฏใช้ (structural preferences) เช่น ศึกษาว่าคำคุณศัพท์นั้นๆมักเกิดในลักษณะที่เป็น attribute หรือ predicate หรือศึกษาว่าคำกริยาตัวนี้เกิดในโครงสร้างแบบสกรรมกริยา (transitive) หรือ อกรรมกริยา(intransitive) เป็นต้น

ในทำนองเดียวกัน การศึกษาในระดับโครงสร้างไวยากรณ์ (grammatical construction) ด้วยคลังข้อมูลภาษาก็สามารถศึกษาในแบบแผนความความสัมพันธ์ระดับคำ (lexical association) และในแบบแผนความสัมพันธ์ระดับไวยากรณ์ (grammatical association) ได้ การศึกษาด้านแบบแผนความสัมพันธ์ในระดับคำ ก็คือการศึกษาการเกิดร่วมกันของโครงสร้างภาษาแบบหนึ่งๆ กับคำบางคำ เช่น ศึกษาว่า that-clause มักเกิดกับคำกริยาหลักประเภทใด หรือศึกษาว่า to-clause เกิดกับคำกริยาหลักใดได้บ้าง เป็นต้น สำหรับการศึกษาด้านแบบแผนความสัมพันธ์ระดับไวยากรณ์ ก็คือการศึกษาว่าโครงสร้างภาษามีความสัมพันธ์กับตัวแปรในบริบท (contextual factors) อย่างไร เช่น ศึกษาการ

ปรากฏของ that-clause ในโครงสร้างแบบ extraposed construction เปรียบเทียบกับที่เกิดใน to-clause เป็นต้น

2. การศึกษาระดับคำ

ในกรณีของการจัดทำพจนานุกรม การใช้คลังข้อมูลภาษาร่วมกับโปรแกรมคอนคอร์แด๊นซ์จะช่วยให้ผู้วิจัยเข้าใจความหมายและการใช้คำๆนั้นได้ดีขึ้นเนื่องจากได้เห็นตัวอย่างมากมาย เห็นความหลากหลายของการใช้คำๆนั้น และช่วยให้ผู้ศึกษาสามารถเข้าใจความหมายต่างๆ ของคำนั้นได้ดีและถูกต้องยิ่งขึ้น ในปัจจุบัน การจัดทำพจนานุกรมนั้นจำเป็นต้องอาศัยคลังข้อมูลภาษาเป็นพื้นฐานในการวิเคราะห์ พจนานุกรมที่จัดทำโดยไม่อาศัยคลังข้อมูลภาษาจะไม่ได้รับการยอมรับและเชื่อถือมากเท่าพจนานุกรมที่จัดทำโดยใช้คลังข้อมูลภาษา ดังจะเห็นได้จากการแข่งขันกันของสำนักพิมพ์ต่างๆในการทำพจนานุกรมโดยอาศัยคลังข้อมูลภาษา เช่น Collins Cobuild English Dictionary ใช้ข้อมูลจาก The Bank of English, Cambridge Dictionary of American English ใช้ข้อมูลจาก Cambridge International Corpus, Longman Dictionary of American English ใช้ข้อมูลจาก The Longman Corpus Network คลังข้อมูลภาษาเหล่านี้มีขนาดมากกว่า 100 ล้านคำขึ้นไป นอกจากนี้ การใช้คลังข้อมูลภาษายังช่วยให้เราได้รายการคำที่เป็นคำเกิดใหม่ (neologism) เช่นที่ Renouf (1993) พบคำเกิดใหม่ 5,374 คำจากคลังข้อมูล London Times ขนาด 2.5 ล้านคำ ได้แก่ *poserphone, wetware, empurple, indifferentness* เป็นต้น ซึ่งการตรวจหาคำเกิดใหม่นี้เป็นงานที่ผู้จัดทำพจนานุกรม (lexicographer) ให้ความสนใจ นอกจากนี้ผู้จัดทำพจนานุกรมยังต้องตัดสินว่าจะกำหนดให้คำใดเป็นรายการคำหลัก (entry) ในพจนานุกรม การศึกษาคำจากคลังข้อมูลจะช่วยให้ตัดสินใจได้อย่างมั่นใจมากขึ้น สำหรับนักภาษาศาสตร์ประยุกต์ที่ศึกษาเรื่องของคำสามารถ

ใช้ประโยชน์จากคลังข้อมูลภาษาในการวิเคราะห์ได้อย่างน้อยใน 3 รูปแบบดังนี้

1. เพื่อศึกษาคำหลายหน้าที่ (multifunctional words) ตัวอย่างเช่น Aijmer (1986) ได้ศึกษาหน้าที่ต่างๆของคำว่า *actually* หรือ Altenberg (1994) ได้ศึกษาหน้าที่ต่างๆของคำว่า *such*

2. เพื่อศึกษาการใช้ของคำบางคำที่มีลักษณะคล้ายคลึงเกี่ยวข้องกัน ตัวอย่างเช่น Kennedy (1991a) ศึกษาความเหมือนและความแตกต่างของการใช้คำว่า *between* และ *through* Collins (1991) ศึกษาเรื่องของ modal verb ต่างๆ

3. เพื่อศึกษาการใช้คำหรือวลีที่มักปรากฏใช้แบบตายตัว (the use of relatively fixed lexical expressions) หรือที่เรียกว่าวลีเสมือนคำ (lexical phrase) ตัวอย่างเช่น การศึกษาเรื่องคำปรากฏร่วม (collocation) ของ Renouf และ Sinclair (1991)

การศึกษาอีกประเภทหนึ่งที่สำคัญ คือ การศึกษาคำในกลุ่มที่มีความหมายคล้ายคลึงกันได้ กรณีที่พจนานุกรมแสดงคำต่างๆ ที่อยู่ในกลุ่มความหมายเดียวกัน การศึกษาผ่านคลังข้อมูลภาษาจะช่วยให้เห็นถึงวิธีที่เจ้าของภาษาใช้คำเหล่านี้ในรูปแบบประโยคที่ต่างๆ กัน ซึ่งความแตกต่างของการใช้ดังกล่าวจะบอกได้ถึงหน้าที่ของคำแต่ละคำ เช่น Biber et al. (1996) ศึกษาคำว่า *big* กับ *large* และ *little* กับ *small* ว่ามีการใช้แตกต่างกันอย่างไร ซึ่งจากการศึกษาโดยใช้คลังข้อมูลภาษาพบว่า *big-little* มักใช้ในภาษาพูดมากกว่า *large-small* ในขณะที่ *large-small* มักใช้ในตัวบททางวิชาการมากกว่า ซึ่งความแตกต่างตรงนี้ หากนำมาวิเคราะห์ศึกษาต่อก็จะพบว่าคำสองคู่นี้มีกลุ่มคำปรากฏร่วมที่แตกต่างกัน (การศึกษาในลักษณะนี้เป็นการศึกษาที่แบบแผนความ

ความสัมพันธ์ระดับคำ) โดยที่ *big-little* มักใช้กับคำนามที่เป็นสิ่งมีตัวตน (concrete) และมักเป็น animate เช่น *boys*, *things*, *girls* ในขณะที่ *large–small* มักจะใช้กับคำนามที่บอกปริมาณ เช่น *amount*, *number*, *proportion* ผลที่ได้นี้ทำให้เราสามารถอธิบายได้ว่าทำไม *large-small* ถึงใช้ในตัวบททางวิชาการมากกว่า ทั้งนี้เพราะเนื้อหาทางวิชาการมักเป็นเรื่องที่เกี่ยวกับการวัดจำนวนปริมาณ และหากศึกษาที่แบบแผนความสัมพันธ์ระดับไวยากรณ์ (grammatical structure association) ก็จะพบว่าคำทั้งสองคู่นี้มีความแตกต่างกัน ถึงแม้ *little-small* จะสามารถใช้ได้ทั้งในลักษณะที่เป็น attributive (*I'm a small person*) หรือเป็น predicative (*It is too small*) *small* จะใช้ในลักษณะของ predicative มากกว่า *little*

3. การศึกษาระดับโครงสร้าง

คลังข้อมูลภาษาสามารถนำมาใช้ในการศึกษาวิจัยเรื่องทางโครงสร้างไวยากรณ์เพื่อค้นหาคำตอบในประเด็นคำถามต่างๆ ดังตัวอย่างที่จะกล่าวถึงต่อไปนี้

- ศึกษาว่าโครงสร้างหนึ่งๆมีการใช้ในลักษณะใดได้บ้างและเมื่อเปรียบเทียบกับโครงสร้างอื่นๆที่เกี่ยวข้องหรือคล้ายคลึงกัน ทั้งสองโครงสร้างมีความเหมือนหรือความแตกต่างกันอย่างไรบ้าง
- ศึกษาเปรียบเทียบกับโครงสร้างที่เกี่ยวข้องหรือคล้ายคลึงดังกล่าวว่ามีปรากฏใช้มากหรือน้อยเพียงใด
- ศึกษาว่าโครงสร้างนั้นๆมีการใช้ตามระดับทำเนียบภาษา (register) ที่ต่างกันอย่างไร มีปริมาณการใช้มากน้อยเพียงใด
- ศึกษาว่าโครงสร้างนั้นๆมักจะปรากฏใช้บ่อยเฉพาะกับคำบางคำหรือไม่
- ศึกษาว่ามีปัจจัยอะไรบ้างในบริบทที่เกี่ยวข้องที่ทำให้มีการใช้รูปแปรต่างๆ

ของโครงสร้างไวยากรณ์

ในปัจจุบัน มีหนังสือและงานวิจัยจำนวนมากที่เกี่ยวข้องกับการศึกษาโครงสร้างไวยากรณ์โดยอาศัยคลังข้อมูลภาษา นอกจากนี้คลังข้อมูลภาษายังสามารถนำมาใช้ในการศึกษาอิทธิพลของบริบทที่มีต่อการใช้โครงสร้างแบบต่างๆ ได้ เช่น Granger (1983) ศึกษาเรื่องของประโยคแบบ passive Meyer (1992) ศึกษาเรื่องของประโยคแบบ apposition Johansson (1995) ศึกษาเปรียบเทียบประพันธสรรพนาม (relativizers) *whose* และ *of which* เป็นต้น

Biber et al. (1996) ยังยกตัวอย่างการศึกษาเปรียบเทียบโครงสร้างประโยคสองแบบที่มีความหมายคล้ายกัน เช่น *I hope that I can go* กับ *I hope to go* ซึ่งจากการศึกษาจากคลังข้อมูลภาษาพบว่า ทั้งสองโครงสร้างนี้มีการใช้ที่ต่างกัน โดย that-clause จะพบมากในภาษาพูด แต่จะไม่ค่อยพบมากในเอกสารทางวิชาการ ซึ่งเมื่อพิจารณาด้านแบบแผนความความสัมพันธ์ระดับคำ (lexical association) ก็พบว่าคำกริยาส่วนใหญ่จะเลือกเกิดกับโครงสร้างใดโครงสร้างหนึ่งเป็นหลัก เช่น *imagine, mention, suggest, conclude* จะใช้กับ that-clause ส่วน *begin, start, like, love, try, want* จะใช้กับ to-clause และเมื่อพิจารณาแบบแผนความสัมพันธ์ระดับไวยากรณ์ (grammatical association) โดยพิจารณาโครงสร้างประโยคแบบ extraposed (เช่น *It's possible that it'll happen again*) กับโครงสร้างประโยคแบบ non-extraposed (เช่น *I think that you might be wrong*) ก็พบว่า to-clause ใช้ในโครงสร้างแบบ extraposed มากกว่า that-clause

4. การศึกษาด้าน ESP และการแปรตามทำเนียบภาษา

โดยทั่วไปงานวิจัยในเรื่องปริเฉจวิเคราะห์ (discourse analysis) และการหาลักษณะเฉพาะทางภาษา (linguistic characteristics) ของตัวบทต่างๆมัก

เป็นการวิจัยเชิงประจักษ์ (empirical) ซึ่งทำโดยการวิเคราะห์ข้อมูลภาษาจากตัวบทต่างๆ จึงกล่าวได้ว่างานวิจัยทางด้านปริเฉจวิเคราะห์เป็นงานที่อาศัยคลังข้อมูลภาษาเป็นพื้นฐาน งานด้านปริเฉจวิเคราะห์ นี้มีปรากฏทั้งในงานวิจัยด้าน ESP (English for Specific Purpose) และในด้านภาษาศาสตร์สังคม ซึ่งที่ผ่านมามักเน้นที่การวิเคราะห์ในทำเนียบภาษา (register) แบบใดแบบหนึ่ง

การนำคลังข้อมูลภาษามาใช้ ช่วยให้เราสามารถศึกษาลักษณะของภาษาในทำเนียบภาษาหนึ่งๆ หรืออาจศึกษาไปถึงความต่างของภาษาในทำเนียบภาษาที่ต่างกัน ซึ่งจะเป็นประโยชน์สำหรับการศึกษา ESP เพราะผลที่ได้จะแสดงลักษณะเฉพาะของทำเนียบภาษาหนึ่งๆ ออกมา ในเชิงที่เปรียบเทียบกับทำเนียบภาษาอื่นๆ Biber (1988) ได้เสนอวิธีการวิเคราะห์แบบหลายมิติ (multi-dimensional) โดยการพยายามหาลักษณ์ทางภาษา (linguistic feature) ต่างๆที่มีแนวโน้มว่าจะเกิดร่วมกันและพบในตัวบทแต่ละประเภท ลักษณ์ทางภาษาที่จัดได้ในแต่ละกลุ่มจะถูกมองรวมเป็นหนึ่งมิติ (dimension) Biber (1988) ได้ทดลองกำหนดมิติสำหรับการวิเคราะห์ไว้ 6 มิติ เช่น มิติที่1 เป็นเรื่องของ Involved และ Informational Production มิติ 5 เป็นเรื่องของ Non-impersonal และ Impersonal Style เป็นต้น ในแต่ละมิติจะมีการกำหนดลักษณ์ที่เป็นลักษณ์ทางบวก (positive feature) และลักษณ์ทางลบ (negative feature) ไว้ เช่น ในมิติที่ 1 (เรื่องของ Involved และ Informational Production) สิ่งที่เป็นลักษณ์ทางบวก ได้แก่ mental verbs, that-deletion, contraction, 2nd person pronouns, demonstrative pronouns, pronoun IT, etc. ส่วนสิ่งที่เป็นลักษณ์ทางลบ ได้แก่ noun, preposition, attributive adjectives, etc. จากนั้นให้มีการคำนวณค่าในแต่ละมิติของตัวบทที่พบในแต่ละทำเนียบภาษา ด้วยวิธีการเช่นนี้ นักวิจัยจะสามารถเปรียบเทียบตัวบทในทำเนียบภาษาต่างๆว่ามีความ

แตกต่างในมิติต่างๆอย่างไรบ้าง เช่น ตัวบทที่เป็นการสนทนาจะมีคะแนนของมิติด้าน involved และ non-impersonal สูง ส่วนตัวบทด้านชีววิทยาจะมีคะแนนหนักไปทางด้าน informational และเป็น impersonal เป็นต้น

ผลจากการคำนวณคะแนนในมิติต่างๆนี้ สามารถนำมาเปรียบเทียบระหว่างทำเนียบภาษาต่างๆ ได้ การศึกษาโดยเปรียบเทียบทำเนียบภาษานั้น ถูกนำมาใช้ในการหาคำตอบในการวิจัยทางภาษาศาสตร์ รวมทั้งสามารถนำมาศึกษาความสัมพันธ์ระหว่างภาษาเขียนและภาษาพูดในภาษาอังกฤษ (Biber 1988) และศึกษาประวัติความเป็นมาของพัฒนาการภาษาเขียน (Atkinson 1992) นอกจากนี้ การเปรียบเทียบตัวบทจากหลายๆมิติ สามารถนำมาอธิบายลักษณะเฉพาะทางภาษาของ ESP และ EAP ได้ ตัวเลขค่าที่คำนวณได้ในแต่ละมิตินี้จะเป็นประโยชน์ต่อการจัดประเภทของตัวบทต่างๆ

5. **หัวข้อในการวิจัยทางภาษาโดยใช้คลังข้อมูล**

นอกจากการมองแนวทางการวิจัยภาษาโดยดูแบบแผนความสัมพันธ์ (association patterns) ตามที่ Biber et al. (1996) ได้กล่าวไว้ Kennedy (1998) ก็ได้สรุปถึงหัวข้อเรื่องต่างๆที่สามารถนำคลังข้อมูลภาษามาใช้เพื่อประโยชน์ในการศึกษาวิจัยทางภาษาได้ โดยได้ยกตัวอย่างงานวิจัยในหัวข้อเรื่องต่างๆที่ได้ทำมาโดยอาศัยคลังข้อมูลภาษา ดังต่อไปนี้

5.1 การศึกษาเรื่องคำ

การศึกษาเรื่องคำเป็นการศึกษาที่คลอบคลุมเรื่องของการศึกษาความถี่ การปรากฏและการเกิดร่วมกันของคำต่างๆดังที่ได้กล่าวมาแล้ว

5.2 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องคำและหน่วยคำ

การศึกษาเรื่องรูปกริยา (verb form) กับ tense และ aspect ต่างๆ

ตัวอย่างเช่น Ota (1963) ศึกษารูปกริยาจากคลังข้อมูลภาษาอังกฤษ 150,000 คำ พบว่ามีคำกริยาที่อยู่ในรูป finite จำนวน17,166 รูป ซึ่งในจำนวนนี้ คำกริยา 977 ตัวถูกใช้เป็น passive voice นอกจากนี้ยังพบว่า *today, this year* ใช้กับคำกริยาที่อยู่ในรูป past tense มากกว่า present tense และคำกริยา 7 ตัว ได้แก่ *be, think, have, know, say, want, go* มีปริมาณเป็นจำนวน 50% ของคำกริยาทั้งหมดในคลังข้อมูล การศึกษาของนักวิจัยคนอื่นๆที่ตามมา เช่น George (1963), Joos (1964) ก็ได้ผลที่คล้ายๆกัน คือพบว่ามีการใช้คำกริยาในรูปของ simple present tense กับ simple past tense มากกว่ารูป tense อื่นๆ โดยที่ simple present จะพบมากในภาษาพูด และ simple past จะพบมากในภาษาเขียน ส่วน perfect กับ progressive aspects นั้นใช้น้อย ผลที่ได้นี้ยังสอดคล้องกับการวิเคราะห์จากคลังข้อมูล Brown ของ Francis และ Kucera (1982)

การศึกษาเรื่อง modal

Coates (1983) ศึกษาเรื่อง modal โดยใช้คลังข้อมูล London-Lund และ LOB ซึ่งจากการศึกษาพบว่ามีความแตกต่างระหว่างการใช้ modal ในภาษาพูดและภาษาเขียน โดยที่ *would* จะใช้มากในภาษาเขียนของภาษาอังกฤษแบบอังกฤษ ในขณะที่ *will* จะใช้มากในภาษาพูดของภาษาอังกฤษแบบอังกฤษ นอกจากนี้ ยังพบความแตกต่างของการใช้ modal ในโครงสร้างกริยาวลีแบบต่างๆ

การศึกษาเรื่อง voice

Svartvik (1966) ศึกษาเรื่องของประโยคแบบ passive ในตัวบทจากทำเนียบภาษาต่างๆ แปดทำเนียบ ผลการศึกษาพบว่าความแตกต่างของการใช้

ขึ้นอยู่กับว่าตัวบทนั้นเป็นตัวบทประเภทข้อมูลข่าวสาร (informative) หรือเป็นประเภทจินตนาการ (imaginative)

การศึกษาการใช้คำกริยาและคำ particle

Akkerman (1984) ศึกษากริยาวลีในคลังข้อมูล Brown ผลการศึกษาพบว่าคำ particle ว่า *up* และ *out* มีจำนวนเป็น 47.4% ของคำ particle ทั้งหมดที่ปรากฏหลังคำกริยา และพบว่า ในกรณีที่มีการใช้ particle 2 ตัวติดกัน particle ตัวที่สองบางตัวก็สามารถละได้ (optional) แต่บางตัวก็จำเป็นต้องมี (obligatory) เช่น ในกริยาวลี *go in for* จำเป็นต้องมี คำว่า *for* ปรากฏอยู่ ในขณะที่ ในกริยาวลี *fall out with* สามารถที่จะละ *with* ได้ นอกจากนี้ยังพบอีกว่า ประมาณ 2 ใน 3 ของ particle ตัวที่สองสามารถละได้ ส่วนในกรณีที่ particle แรกเป็น *around, down, away* พบว่ามีแนวโน้มที่จะบังคับให้ต้องมี particle ตัวที่สอง

การศึกษาเรื่อง subjunctive

Johansson และ Norheim (1988) ได้ศึกษา subjunctive เพื่อเปรียบเทียบความต่างระหว่างภาษาอังกฤษแบบอังกฤษและภาษาอังกฤษแบบอเมริกัน โดยศึกษาจากคลังข้อมูล Brown และ LOB ผลการศึกษาพบว่า mandative subjunctive พบมากในภาษาอังกฤษแบบอเมริกัน ตัวอย่างเช่น *It is important that he join us* ในขณะที่ภาษาอังกฤษแบบอังกฤษจะนิยมใช้ *should* มากกว่า

การศึกษาเรื่องคำบุพบท

Mindt และ Weber (1989) ศึกษาคำบุพบท พบว่า 12.21% ของคำทั้งหมดในคลังข้อมูลจะเป็นคำบุพบท โดยมี *of* คิดเป็น 30% ของคำบุพบททั้งหมด Sinclair (1991) ศึกษาคำว่า *of* พบว่า *of* นั้นขึ้นกับคำที่มาปรากฏ

ข้างหน้ามากกว่าคำข้างหลัง (มีความสัมพันธ์กับคำข้างหน้ามากกว่าคำตามหลัง) ดังนั้น *of* จึงไม่ได้ใช้เหมือนคำบุพบททั่วไป(ที่มีความสัมพันธ์กับคำตามหลัง)

การศึกษาเรื่องคำสันธาน

ตัวอย่างเช่น Kennedy (1998) พบว่า *since* เป็นได้ทั้งคำบุพบทและสันธาน แต่ *since* ใช้เป็นสันธานมากกว่าบุพบท นอกจากนี้ *since* ใช้ได้ทั้งที่เป็นการบอกเวลา (temporal) หรือบอกสาเหตุ (causal) แต่ในตัวบททางวิชาการจะพบการใช้ *since* ที่เป็นการบอกสาเหตุมากกว่า

5.3 การศึกษาเรื่องไวยากรณ์ที่เน้นเรื่องประโยค

5.3.1 การศึกษาความยาวประโยค

Francis และ Kucera (1982) ศึกษาเรื่องของความยาวประโยคจากคลังข้อมูลภาษา พบว่าความยาวประโยคมีความสัมพันธ์กับประเภท (genre) ของตัวบท และตัวบทที่เป็นข้อมูลข่าวสาร (informative) จะมีความยาวประโยคเฉลี่ยมากกว่าตัวบทประเภทจินตนาการ (imaginative) ทั้งนี้เป็นผลมาจากการที่ตัวบทประเภทข้อมูลข่าวสารจะมีการใช้นามวลีแบบซับซ้อน (complex NP) และมีการใช้ภาคแสดงแบบ non-finite มากกว่า

5.3.2 การศึกษากระบวนการทางไวยากรณ์ (syntactic process)

การศึกษาเรื่อง clause pattern

Altenberg (1993) ศึกษา การใช้คำกริยาร่วมกับหน่วยเติมเต็ม (verb-complement combination) มากกว่า 5,000 ตัวอย่าง เพื่อดูว่าลักษณะทางไวยากรณ์แบบไหนของการใช้คำกริยาร่วมกับหน่วยเติมเต็มที่มีการใช้มากที่สุด ผลการศึกษาพบว่าการใช้คำกริยาร่วมกับหน่วยเติมเต็มประเภท copular และ

monostrative จะมีการใช้มากที่สุด ซึ่งสอดคล้องกับที่ Ota (1963) พบว่า be เป็นคำกริยาที่มีปรมาณการใช้มากที่สุด

การศึกษาเรื่องการขยายคำนาม (noun modification)

De Haan (1989) ศึกษาเรื่อง postmodification ในภาษาอังกฤษ จากคลังข้อมูล Nijmegen ผลการศึกษาพบว่านามวลีที่มี postmodification มักจะไม่ใช้ทำหน้าที่ประธาน แต่หากว่าใช้เป็นประธาน ก็มักจะมีแนวโน้มว่าจะใช้เป็น non-restrictive หรือ appositional มากกว่าที่จะเป็น restrictive De Haan (1991) ยังพบว่า 70% ของ relative clause แบบ postmodifying ในข้อมูลเป็น restrictive relative clause และจะไม่พบ zero หรือ *that* กับ relative clauses แบบ non-restrictive

การศึกษาเรื่อง conditionality

ตามที่ Hill (1960) กล่าวว่าในประโยคที่มีคำว่า *if* นั้นสามารถมีการผสมของ finite verb ใน main clause กับ subordinate clause ที่แตกต่างกันได้อย่างน้อย 324 แบบ แต่จากการศึกษาคลังข้อมูล LOB ของ Wang (1991) พบว่ามีรูปแบบที่พบเพียง 76 แบบ และในจำนวนแปดแบบที่ใช้นั้นคิดเป็นจำนวน 50% ของทั้งหมด

การศึกษาเรื่องการปฏิเสธ (negation)

Tottie (1991) ศึกษาเรื่องของการปฏิเสธไว้อย่างละเอียดทั้งในแง่มุมของ grammatical, semantics และ pragmatics โดยศึกษาจากคลังข้อมูล LOB และ SEU ผลการศึกษาพบว่าการปฏิเสธเกิดในภาษาพูดมากกว่าในภาษาเขียนถึงสองเท่า และพบว่ามีนัยยะสำคัญในความต่างระหว่างภาษาพูดและภาษาเขียนในการเลือกใช้ *not* และ *no* กล่าวคือ ในขณะที่ภาษาพูดเลือกใช้ *not* ถึง 66% ภาษาเขียนจะเลือกใช้ *not* เพียง 37%

5.4 การศึกษาวัจนกรรมในภาษาพูด

ตัวอย่างเช่น Svartvik (1992) ศึกษาคำเสริมในปริเฉจ (discourse particle) *well* ซึ่งเป็นคำที่พบมากเป็นลำดับที่ 14 ในการสนทนา ซึ่ง *well* ในภาษาพูดจะไม่ได้ใช้ในความหมายเดียวกับ *well* ในภาษาเขียน *well* ในภาษาพูดใช้เพื่อแสดงการเปลี่ยนหัวข้อ (shift topic) ได้ Aijmer (1996) ศึกษาคำ *sort of* กับ *kind of* ในลักษณะที่เป็นการแสดงความไม่แน่ใจ (hedging device) โดยศึกษาจากคลังข้อมูล London-Lund สองคำนี้เป็นคำที่ใช้แสดงความไม่แน่ใจที่พบมากที่สุด และพบว่า ถึงแม้ *sort of* จะปรากฏหน้าคำนาม คำกริยา คำคุณศัพท์ หรือคำวิเศษณ์ก็ได้ แต่จะปรากฏเป็นคำขยายคำนามมากที่สุด คือ 62% *sort of* กับ *kind of* ยังใช้เพื่อช่วยลดระดับความเข้มข้น (soften the intensity of discourse) และลดความเป็นทางการ (promote informality) เป็นการบอกว่าผู้พูดแค่ประมาณการเอา

5.5 การศึกษาด้านการแปรของภาษา

ตัวอย่างงานวิจัยหลายตัวอย่างที่กล่าวมาแล้วก็จัดอยู่ในประเภทนี้ได้ เช่น งานที่มีการศึกษาเปรียบเทียบระหว่างภาษาพูดและภาษาเขียน เช่น การศึกษาคำว่า *such* ของ Altenberg (1994) ในภาษาพูดและภาษาเขียน การศึกษาคำว่า *whose* กับ *of which* ในภาษาพูดและภาษาเขียนของ Johansson (1993) เป็นต้น

นอกจากการศึกษาโดยแยกประเภทตัวบทตามลักษณะภาษาพูดหรือภาษาเขียนแล้ว Biber (1988) ได้เสนอวิธีการแยกประเภทของตัวบทตามลักษณะมิติ (dimension) ดังที่ได้กล่าวมาแล้ว โดยดูลักษณ์ทางภาษาที่ปรากฏในตัวบทนั้นเป็นหลัก ตัวบทประเภทเดียวกันแต่ถ้ามีความแตกต่างของลักษณ์ทางภาษาก็จะจัดเป็นคนละกลุ่ม ส่วนตัวบทต่างประเภทกัน แต่ถ้ามีลักษณ์ทางภาษา

คล้ายคลึงกันก็จะจัดเป็นตัวบทกลุ่มเดียวกัน (เวลาคำนวณให้คำนวณคะแนนของแต่ละลักษณ์ในมิตินั้น ซึ่งมีทั้งค่าบวกและค่าลบ)

การศึกษาการแปรของภาษาอังกฤษในภูมิภาคต่างๆก็จัดว่าเป็นอีกแขนงหนึ่งของการใช้ประโยชน์จากคลังข้อมูลภาษา โครงการ ICE (International Corpus of English) มีขึ้นเพื่อศึกษาเปรียบเทียบเรื่องนี้โดยเฉพาะ หรืองานวิจัยที่เปรียบเทียบระหว่างภาษาอังกฤษแบบอังกฤษและภาษาอังกฤษแบบอเมริกันก็จัดอยู่ในงานประเภทนี้ เช่น ในการเปรียบเทียบคลังข้อมูล LOB กับ Brown (Johansson 1980) พบว่าคำที่ปรากฏมาก 50 คำแรกจะเหมือนกัน จะต่างก็เพียงสองคำ คำว่า *so* ได้ที่ 52 ในคลังข้อมูล Brown แต่ได้ที่ 46 ใน LOB คำว่า *more* ได้ที่ 48 ในคลังข้อมูล Brown แต่ได้ที่ 51 ใน LOB สิ่งที่น่าสนใจคือคำที่มีความถี่ต่ำทั้งหลาย เพราะความแตกต่างของคำเหล่านี้สามารถสะท้อนให้เห็นความแตกต่างระหว่างสองวัฒนธรรมได้ Johansson (1980) เสนอวิธีคำนวณเทียบระหว่าง LOB กับ Brown จากสูตรนี้

(ความถี่ที่พบใน LOB - ความถี่ที่พบใน Brown) / (ความถี่ที่พบใน LOB + ความถี่ที่พบใน Brown)

คือ นำเอาความถี่ของคำนั้นที่พบในคลังข้อมูล LOB และในคลังข้อมูล Brown มาคำนวณ ค่าที่คำนวณได้นี้จะอยู่ระหว่าง -1 ถึง 1 ถ้าได้ค่าที่เป็นบวกแสดงว่าคำนั้นมีความโน้มเอียงไปทาง LOB หรือเป็นคำที่มีลักษณะเป็นภาษาอังกฤษแบบอังกฤษมากกว่า ถ้าเป็นลบก็มีความโน้มเอียงไปทาง Brown หรือเป็นคำที่มีลักษณะเป็นภาษาอังกฤษแบบอเมริกันมากกว่า ด้วยวิธีนี้ก็จะได้คำที่ถือเป็นตัวแทนของภาษาอังกฤษในแต่ละแบบได้ ดังนี้

```
+0.50 to +1.0
accommodation, acquaintances, advertisement, afterwards,
alderman, amongst, barister(s), beforehand, booklet,
borough(s), borstal, castle, chap(s), cheque, cinema(s),
```

commonwealth, constable, constituency, councillor(s), criket, cupboard, draught, estates, excellency, exchequer, exhaust, expenditure, firstly, flats, forecast(s), fortnight, gaol, gentry, grey, grid, guinea(s), hairdresser, hitherto, holiday, icing, incomes, jug(s), keen, kin, lads, lodgings, lorry, ltd., madam, maiden, minister, ministry, monarch(y), mum, owing, packet, parish, parliament(ary), petrol, premises, prisoner, railway(s), registrar, rightly, ruler, scales, scheme(s), secondly, shareholders, shilling, sir, specially, spectacles, tea, thirdly, turnover, vicar, ware, whence, whilst, withdrawal, yacht

คำที่เป็นภาษาอังกฤษแบบอังกฤษ

-0.50 to -1.00
accommodations, afterward, ain't, apartment(s), assignment(s), attorney, auto, automobile(s), barbecue, baseball, basement, billion(s), boost, bulletin, businessman, cafeteria, calendar, campus, candy, carryover, cents, chores, closet, commerce, confederate, congress(ional), congressman, cop(s), cowboy, currently, dealer(s), desegragation, detergent, determine, dollar(s), downtown, driveway, elevator, ethnic, fallout, fled, formulas, freight, gasoline, gonna, gotten, governor, graduate, grandma, gray, guy, highway(s), jazz, Jr., kid, legislature, lest, liquor, lobby, locates, location(s), lots, lumber, missile, movie(s), Negro, network, onto, percent, phoney, prairie, railroad(s), ranch, realtors, Republican(s), salesman, salesmen, sauce, shear, sidewalk, someday, sponsor, streetcar, therapist, trailer(s), transportation, truck(s), turnpike, unwed, uptown, utility, vacation(s), veterans, wagon(s), wildlife, yell.

คำที่เป็นภาษาอังกฤษแบบอเมริกัน

(Kennedy, 1998: 192)

นอกจากนี้ การศึกษาการแปร (variation) ระหว่างทำเนียบภาษาก็เป็นอีกเรื่องหนึ่งที่สามารถใช้ประโยชน์จากคลังข้อมูลภาษาได้ และเรื่องการ

เปรียบเทียบภาษาในเชิงประวัติ (diachronic) ก็สามารถนำคลังข้อมูลภาษามาใช้ประโยชน์ได้ โดยทำการเปรียบเทียบจากคลังข้อมูลที่เป็นตัวแทนภาษาในช่วงเวลาต่าง ๆ กัน

5.6 การศึกษาด้านวัจนลีลา

การศึกษาวัจนลีลา (stylistics) เป็นการศึกษาที่มีมาแต่เดิม เป็นการศึกษาวัจนลีลาการเขียนในเชิงวรรณกรรมของนักเขียนว่ามีลักษณะเด่นแตกต่างจากการใช้ภาษาทั่วไปอย่างไร งานการศึกษาวัจนลีลาโดยเนื้องานแล้วมีความสอดคล้องกับการศึกษาทางภาษาศาสตร์คลังข้อมูล เพราะทั้งสองฝ่ายต่างก็เป็นการศึกษาแบบเชิงประจักษ์ (empirical) ต้องอาศัยข้อมูลภาษาที่เกิดขึ้นเป็นพื้นฐานการวิเคราะห์ มีการมองข้อมูลทั้งในเชิงปริมาณและเชิงคุณภาพโดยการเปรียบเทียบตัวบทที่วิเคราะห์กับตัวบทอื่น

Wynne (2005) สรุปภาพรวมการศึกษาวัจลลีลาโดยอาศัยคลังข้อมูลภาษาว่าจำแนกได้เป็นสองแนวทาง แนวทางแรกใช้วิธีวิเคราะห์และกำกับผลลงบนตัวบทที่วิเคราะห์ (corpus annotation) จากนั้นจึงจะสามารถสรุปภาพรวมการวิเคราะห์ออกมา ซึ่งเมื่อก่อนใช้วิธีการนับเอาเองแต่ในปัจจุบันสามารถใช้คอมพิวเตอร์ช่วยแจกแจงผลออกมาได้ แนวทางที่สองเป็นการศึกษาตัวบทนั้นว่ามีความเบี่ยงเบนออกจากมาตรฐานอย่างไร ทำไมงานเขียนนั้นจึงมีลักษณะการใช้ภาษาที่แปลกเด่นจากงานทั่วๆไป โดยมองในเรื่องของคำปรากฏร่วมที่ใช้ (collocation) เรื่องของหมวดคำปรากฏร่วม (colligation) และเรื่องเฉดความหมาย (semantic prosody) เพราะเป็นการมองหารูปแบบการใช้คำหรือวลีที่สนใจว่า มีการใช้ร่วมกับคำกลุ่มใดเป็นพิเศษ หรือทำให้เห็นรูปแบบหมวดคำที่ปรากฏร่วมกับคำหรือวลีที่สนใจ หรือกลุ่มคำที่ใช้ร่วมกับคำที่สนใจนั้นมักมีเฉดความหมายไปในทำนองใด ความคิดในเรื่องการวิเคราะห์คำหรือวลีในลักษณะนี้

สะท้อนให้เห็นว่าความหมายไม่ใช่เรื่องเฉพาะของคำหรือวลีนั้น แต่เป็นเรื่องที่เกี่ยวข้องกับคำอื่นๆที่ปรากฏร่วมด้วย แนวคิดนี้ Hoey (2005) ได้นำเสนอเป็นทฤษฎีเรื่อง lexical priming ซึ่งอธิบายความหมายของคำโดยอิงแนวคิดของภาษาศาสตร์ปริชานว่า คนเราเข้าใจความหมายของคำ ไม่ใช่จากความหมายที่ใส่อยู่ในคำแบบที่เชื่อกัน แต่เป็นการประกอบกันของความหมายต่างๆ ของคำนั้นกับบริบทที่มักจะปรากฏใช้ด้วยเป็นปกติวิสัย งานเขียนที่สร้างสรรค์คืองานที่มีการใช้คำซึ่งไม่เข้ากับครรลองที่คำนั้นปกติจะใช้กันอยู่ จึงเกิดเป็นความแปลกเด่นขึ้นได้

อย่างไรก็ตาม Wynne ตั้งข้อสังเกตว่า แม้งานด้านวัจนวีลาจะมีเนื้อหาและวิธีวิทยาที่เข้ากันกับการศึกษาทางด้ายภาษาศาสตร์คลังข้อมูล แต่งานด้านวัจนลีลาที่ใช้คอมพิวเตอร์ช่วยในการศึกษาอย่างเต็มรูปแบบที่ผ่านมากลับมีไม่มากเท่าที่ควร เป็นไปได้ว่าเพราะผู้ที่ทำงานด้านการศึกษาวัจนลีลานี้เป็นพวกที่ศึกษาทางวรรณกรรมเป็นหลัก ไม่มีถนัดในการใช้เครื่องมือทางคอมพิวเตอร์ การกำกับข้อมูลภาษาตามมาตรฐานภาษาศาสตร์คลังข้อมูล หรือที่ไม่อยากใช้คอมพิวเตอร์ ส่วนหนึ่งอาจเชื่อว่าการใช้คอมพิวเตอร์เป็นเครื่องมือการวิเคราะห์เป็นการทำงานแบบวิทยาศาสตร์ซึ่งไม่ใช่มนุษยศาสตร์แบบที่พวกตนเองเป็นอยู่

บทที่ 4 คลังข้อมูลกับการแปล

1. บทนำ

ในการแปลเอกสารฉบับหนึ่ง นอกจากนักแปลจะใช้ทรัพยากรการแปลทั่วๆไป เช่น พจนานุกรมทั่วไป พจนานุกรมเฉพาะทาง พจนานุกรมสองภาษา สารานุกรมเรื่องต่างๆ ฯลฯ เพื่อประโยชน์ในการหาคำอธิบายศัพท์หรือคำแปลหรือทำความเข้าใจตัวบทที่ต้องการแปลแล้ว นักแปลยังสามารถใช้คลังข้อมูลภาษาและโปรแกรมอย่างเช่น โปรแกรมคอนคอร์แด็นซ์ ให้เป็นประโยชน์ต่อการแปลโดยเฉพาะอย่างยิ่งเพื่อเพิ่มประสิทธิภาพของการแปลให้ดีมากขึ้น เพราะทำให้ผู้แปลได้เห็นตัวอย่างการใช้ที่เป็นจริงในเอกสารแต่ละด้าน นอกจากนั้นการเก็บรวบรวมตัวบทจากการแปลเป็นคลังข้อมูลเทียบบท (parallel corpora) โดยเฉพาะก็ทำให้เราสามารถศึกษาลักษณะเฉพาะต่างๆของภาษาแปลได้ ประโยชน์ของคลังข้อมูลภาษากับการแปลแบ่งออกเป็นสองส่วนใหญ่ๆ ส่วนแรกเป็นประโยชน์ต่อการศึกษาการแปล (translation study) ส่วนที่สองเป็นประโยชน์ต่อการแปลของนักแปลโดยตรง

งานแปลจัดเป็นงานที่มีความสำคัญอย่างยิ่งงานหนึ่งในสังคม เพราะทำให้เราได้รู้จักวรรณกรรมดีๆที่เขียนในภาษาต่างๆ ทำให้ได้รู้เรื่องราวต่างๆไม่ว่าจะเป็นเรื่องทางศิลปะ ทางสังคมวิทยา ทางการเมือง ฯลฯ ซึ่งเราคงไม่สามารถปฏิเสธได้ว่า เราได้ความรู้มาจากหนังสือต่างๆที่มีคนแปลไว้เป็นจำนวนไม่น้อย เช่น หลายๆคนรู้จักงานของดอสโตเยฟสกี้ (Dostoyevsky) ได้อ่านหนังสือเรื่อง “พี่น้องคารามาซอฟ” ถึงแม้ว่าจะไม่สามารถอ่านภาษารัสเซียได้ ก็เพราะมีผู้แปล

หนังสือเล่มนี้ออกเป็นภาษาต่างๆ รวมทั้งภาษาไทยด้วย แต่กระนั้น งานแปลก็ยังเป็นงานที่ไม่ได้รับความสำคัญมากนัก เพราะถูกมองว่างานแปลเป็นงานที่เบี่ยงเบน (distort) ไปจากภาษาเดิม ไม่ควรค่าต่อการศึกษาภาษาอย่างจริงจัง หากจะมีการศึกษาก็มักจะเป็นไปในเชิงเปรียบเทียบว่าแตกต่างหรือเบี่ยงเบนไปจากต้นฉบับของเดิมมากน้อยเพียงใด หรือแม้แต่ในวงการภาษาศาสตร์คลังข้อมูลเอง ข้อมูลภาษาแปลก็ยังถูกกันออกไป ไม่นำมารวมเข้าไว้ในคลังข้อมูลภาษาของหลายๆที่ในยุโรป เช่น คลังข้อมูล BNC เพราะมองว่าข้อมูลภาษาแปลไม่ใช่ตัวแทนข้อมูลที่เป็นภาษาที่แท้จริงของภาษานั้น

แต่ Baker (1993) ได้เสนอมุมมองใหม่ว่า ข้อมูลภาษาแปลนั้นก็เป็นข้อมูลภาษาแบบหนึ่งที่ใช้กันจริง เป็นการใช้ภาษาเพื่อการติดต่อสื่อสาร (communicative event) แบบหนึ่ง ไม่ควรจะมีความต่ำต้อยหรือเหนือกว่าการใช้ภาษาในลักษณะอื่นๆ แน่นอนว่าภาษาที่ได้จากการแปลย่อมมีความแตกต่างจากภาษาแบบอื่นๆ แต่ Baker มองว่าความแตกต่างนี้เป็นเรื่องที่ควรศึกษา และควรเป็นที่สนใจศึกษาของศาสตร์อื่นๆที่เกี่ยวข้องด้วย เช่น ภาษาศาสตร์ วรรณคดี และ การศึกษาการสื่อสารเชิงวัฒนธรรม (cultural communication studies) ยิ่งไปกว่านั้น ควรจะมีศาสตร์ที่มองเรื่องการแปลเป็นศูนย์กลางและเป็นหัวใจของการศึกษานั้น ซึ่งปัจจุบันก็มีเกิดขึ้นบ้างแล้วและรู้จักกันในชื่อของศาสตร์การแปล (science of translation หรือ translation studies)

Baker กล่าวว่า เราจะสามารถเข้าใจพัฒนาการของศาสตร์การแปลได้ดีขึ้น หากพิจารณาความแตกต่างระหว่าง “ศาสตร์” (discipline) กับ “สาขาศึกษา” (field of study) ตามที่ Eco (1976) ได้เสนอไว้ โดย Eco อธิบายว่า “ศาสตร์” หมายถึงการมีแนวทางการศึกษาที่มีวิธีการศึกษาและมีสิ่งที่ต้องการศึกษาชัดเจน (its own method and a precise object) ในขณะที่ “สาขาศึกษา” เป็น

การศึกษาที่ยังไม่มีแนวทางการศึกษาที่ชัดเจนเป็นรูปเป็นร่าง (not yet completely unified) ถ้ามองตามหลักที่ว่านี้ ศาสตร์การแปลในช่วงแรกก็จัดอยู่ในประเภทหลังนี้ เพราะยังคงติดอยู่กับการให้ความสำคัญกับความสัมพันธ์ระหว่างต้นฉบับและบทแปลมากกว่าที่จะสนใจในธรรมชาติของตัวภาษาแปลเอง จุดนี้สะท้อนให้เห็นถึงความล้มเหลวที่จะพัฒนาทฤษฎีหรือ “ศาสตร์” ที่จะกำหนดสิ่งที่ต้องศึกษา (object of study) ให้ชัดเจน คือแทนที่จะศึกษาลักษณ์ต่างๆ (feature) ที่ปรากฏในภาษาแปลในฐานะที่เป็นสิ่งที่สามารถศึกษาได้ กลับไปเน้นที่ประเด็นเรื่องการยอมรับ (justify) หรือไม่ยอมรับ (dismiss) บทแปลโดยการนำไปเทียบกับต้นฉบับ

อย่างไรก็ตาม Baker เชื่อว่าแนวโน้มของศาสตร์การแปลในระยะหลังนี้กำลังเปลี่ยนแปลงไป นักวิชาการด้านการแปลได้พยายามกำหนดขอบเขตและเป้าหมายของศาสตร์การแปลใหม่ ซึ่งทั้งนี้เป็นผลมาจากการมีคลังข้อมูลของภาษาต้นฉบับและภาษาแปลมากขึ้น การมีคลังข้อมูลของภาษาแปลที่มีขนาดใหญ่มากขึ้นจะช่วยให้เรามองเห็นและศึกษาบทแปลได้ชัดมากขึ้นว่า การแปลคืออะไรและมีลักษณะเฉพาะตัวอย่างไร

2. ความเป็นมาของศาสตร์การแปล

ในเรื่องของศาสตร์การแปลที่ผ่านมานั้น มีความเชื่อเบื้องต้น (assumption) 2 ประการที่แฝงอยู่ ประการแรก คือ การยึดต้นฉบับเป็นสำคัญ (primacy of the source text) โดยคิดว่าผู้แปลจะต้องแปลให้ถูกต้องและซื่อตรง (faithful) ต่อตัวต้นฉบับ ประการที่สองเป็นผลสืบเนื่องจากความเชื่อเบื้องต้นอย่างแรก คือ การเทียบเท่า (equivalence) ต้นฉบับ คือมองว่าตัวบทแปลจะต้องเหมือนกับต้นฉบับเดิมให้มากที่สุด ทั้งด้านความหมายและลักษณะทางโครงสร้างหรือจำนวนหน่วยภาษา (element) สิ่งเหล่านี้สะท้อนให้เห็นว่า ที่ผ่านมาไม่มี

การมองว่าการแปลเป็นปรากฏการณ์หนึ่งของการใช้ภาษาที่สามารถนำมาศึกษาได้ แต่ไปมุ่งที่จะหาสิ่งที่จะเป็นการแปลที่สมบูรณ์แบบที่จะไม่ทำให้เบี่ยงเบนไปจากต้นฉบับ

แต่ต่อมา ก็เริ่มมีผู้มองเห็นปัญหาของการแปลกันมากขึ้น เริ่มมีการยอมรับกันว่าการแปลที่สมบูรณ์แบบนั้นเป็นสิ่งที่เป็นไปไม่ได้ และได้มีการเปลี่ยนแนวทางจากการหาทางที่จะทำการแปลอย่างสมบูรณ์ (ideal translation) มาสนใจว่าอะไรบ้างที่จะสามารถทำให้เทียบเท่าต้นฉบับได้และอยู่ภายใต้เงื่อนไขใด แต่สิ่งนี้ก็ยังสะท้อนให้เห็นถึงแนวคิดเดิมของการให้ความสำคัญกับความเชื่อเบื้องต้นแต่เดิมที่มองว่าต้นฉบับนั้นสำคัญกว่าบทแปล

นอกจากนี้ แต่เดิมยังมีการมองว่า การแปลคือการถ่ายทอดความหมาย (transfer of meaning) จากต้นฉบับไปยังบทแปล ดังจะเห็นได้จากการให้ความสำคัญกับการแปลแบบคำต่อคำ (word-for-word translation) และการแปลแบบความหมายต่อความหมาย (sense-for-sense translation) แต่ปัจจุบันแนวคิดดังกล่าวเริ่มเปลี่ยนไป และมีแนวคิดใหม่เกิดขึ้น โดย Firth (1968) เสนอให้พิจารณาความหมายของประโยคในลักษณะที่อิงกับบริบทแวดล้อม (situation) แทนที่จะพิจารณาความหมายของประโยคจากความหมายของคำ หรือคือการพยายามเชื่อมโยงระบบและโครงสร้างของภาษาเข้ากับระบบและโครงสร้างทางบริบทแวดล้อมมากกว่าที่จะเชื่อมโยงไปที่ระบบความหมายโดยตรง (connect structure and system of language to structures and systems in the context of situation (as opposed to structures and systems of thought))

ในทำนองเดียวกัน Haas (1968) ก็มองว่า การเทียบเท่าทางความหมายระหว่างต้นฉบับและบทแปล คือการที่มีการใช้ที่สอดคล้องกัน (correspondence between their uses) การเปลี่ยนแปลงในการมองเรื่อง

ความหมายดังกล่าวนับเป็นการเปลี่ยนแปลงจากการให้ความสำคัญจากเรื่องมโนทัศน์ (concept) มาเป็นการให้ความสำคัญกับบริบทการใช้ (situational perspective) หรือเป็นการเปลี่ยนจากความหมายสู่การใช้ ซึ่งการศึกษาเรื่องการใช้ภาษา (usage) จะเป็นไปได้ก็ต้องอาศัยการศึกษาจากข้อมูลจริงที่มีจำนวนที่มากพอ นอกจากนี้ ยังเกิดแนวคิดทฤษฎีบางอย่างขึ้น ซึ่งช่วยส่งผลให้ความเชื่อเบื้องต้นที่ว่าบทแปลจะต้องเหมือนกันกับต้นฉบับนั้นลดความสำคัญลงไป ทฤษฎีที่ว่านี้ คือ ทฤษฎีระบบหลายขั้ว (polysystem theory) และแนวคิดเรื่องรูปแบบปกติ (norm)

3. ทฤษฎีระบบหลายขั้ว

Even-Zohar (1990) ได้เสนอทฤษฎีทางงานวรรณกรรม (literature) ในลักษณะที่เป็นระบบหลายขั้ว (polysystem) คือแทนที่จะมองงานวรรณกรรมว่าเป็นเพียงการรวบรวมตัวบทต่างๆ (collection of texts) เขาเสนอให้มองงานวรรณกรรมในลักษณะที่มีลำดับชั้น (hierarchy) ที่เปลี่ยนแปลงได้ โดยที่ระบบหลายขั้วทางวรรณกรรม (literature polysystem) นี้จะเป็นส่วนหนึ่งของระบบหลายขั้วที่ใหญ่กว่า คือ ระบบหลายขั้วทางวัฒนธรรม (cultural polysystem) ซึ่งก็จะแตกต่างกันไปตามแต่ละวัฒนธรรม ในมุมมองของทฤษฎีนี้ ต้นฉบับและบทแปลเป็นตัวบทที่อยู่คนละระบบ (system) แต่มีความสัมพันธ์ต่อกัน ในมุมมองนี้ บทแปลจึงมีความสำคัญในตัวเองในฐานะที่เป็นระบบหนึ่งในระบบหลายขั้วทางวรรณกรรม

กล่าวคือ ทฤษฎีนี้มองว่ามีระบบ (system) และระบบย่อย (subsystem) มากมายอยู่ภายในระบบหลายขั้วนี้ ในมุมมองนี้ วรรณกรรมแปล (translated literature) ก็เป็นระบบย่อยหนึ่งที่มีความสัมพันธ์กับระบบย่อยของวรรณกรรมต้นฉบับ (source text) แต่ก็มีความเป็นอิสระในตัวเองที่สามารถศึกษาได้

ความสำคัญตรงนี้อาจเปรียบได้กับเหตุการณ์เมื่อครั้งที่ Saussure (1959) ทำให้นักภาษาศาสตร์สามารถกำหนดสิ่งที่นักภาษาศาสตร์ควรสนใจศึกษาได้ คือ Saussure ได้ทำให้นักภาษาศาสตร์เปลี่ยนความสนใจจากเดิมที่สนใจศึกษาเฉพาะเรื่องประวัติความเป็นมาของคำแต่ละคำหรือเปรียบเทียบสืบประวัติความสัมพันธ์ระหว่างภาษาต่างๆ แล้วหันมาให้ความสำคัญกับการศึกษาตัวภาษาโดยตรงเพื่อหาระบบของภาษาขึ้นมา ผลจากการเปลี่ยนแปลงความคิดครั้งนั้น ทำให้สาขาทางภาษาศาสตร์เจริญเติบโตเป็นศาสตร์ขึ้น สิ่งนี้ก็อาจจะเกิดขึ้นกับศาสตร์การแปลได้เช่นกัน ถ้าเราสามารถละความสนใจจากความเชื่อเบื้องต้นเดิมที่มองงานแปลเป็นชิ้นๆไปได้ แล้วหันมาให้ความสำคัญกับการแปลในฐานะที่เป็นระบบหนึ่งในตัวเอง สามารถกำหนดสิ่งที่ต้องการศึกษาได้ชัดเจน ศาสตร์ด้านการแปลก็จะสามารถเจริญเติบโตขึ้นได้ สิ่งที่จะเปลี่ยนไปจากผลอันนี้ คือ การศึกษาเรื่องของการแปลเป็นชิ้นๆเป็นเรื่องๆ จะไม่สำคัญอีกต่อไป แต่การศึกษาบทแปลจำนวนมากๆเพื่อหาลักษณ์ของระบบ (systemic feature) ในภาษาแปลจะเป็นที่สนใจศึกษามากขึ้น

นอกจากนี้ ในระบบหลายขั้ว จะถือว่าแต่ละหน่วย (element) และแต่ละระบบมีลักษณะที่ไม่หยุดนิ่ง ต่างก็จะแข่งขันกันเพื่อแย่งกันเป็นศูนย์กลาง (center) ของระบบมากกว่าที่จะอยู่ชายขอบของระบบ (peripheral system) ทำให้มองได้ว่าวรรณกรรมแปลก็อาจจะเลื่อนเข้ามาเป็นศูนย์กลางและเป็นที่ยอมรับว่าเป็นภาษามาตรฐานได้ และจากการที่ทฤษฎีนี้มองว่าความสัมพันธ์สามารถเกิดได้ทั้งภายในระบบและระหว่างระบบ จึงมีการแทรกแซง (interference) กันระหว่างระบบได้ นอกจากนี้นักทฤษฎีกลุ่มนี้ก็มองว่าบทแปลไม่ได้เป็นเพียงงานที่แปลงจากต้นฉบับเท่านั้น แต่เป็นงานที่สร้างสรรค์ซึ่งผู้แปลสามารถตีความและแปลงให้เข้ากับวัฒนธรรมความเชื่อในภาษาของผู้แปลเองด้วย

ความคิดแบบนี้เอง ที่ทำให้ความสำคัญของต้นฉบับเริ่มลดบทบาทลง Toury (1985) กล่าวว่า การที่การแปลสามารถมีระบบ (system) ของตัวเอง ซึ่งก็คือระบบของภาษาแปล ตัวต้นฉบับก็จะลดความสำคัญลงเป็นเพียงแค่สิ่งเร้า กลายเป็นจุดเริ่มของข้อมูล (source of information) มากกว่าที่จะเป็นจุดตั้งต้นของการวิเคราะห์ กล่าวคือ แม้ว่าคนทุกคนในสังคมนั้นจะรู้ว่ามีต้นฉบับของงานชิ้นนั้นเขียนในภาษาอื่น ตัวต้นฉบับนั้นก็ไม่ได้มีความสำคัญใดใดต่อภาษาในสังคมนั้นๆ แต่ตัวบทที่แปลมาแล้วต่างหากที่จะมีผลกระทบต่อภาษาและวัฒนธรรมของสังคมนั้น

4. แนวคิดเรื่องรูปแบบปกติ

Toury (1978,1980) ได้เสนอแบบจำลองสามส่วน (tripartite model) โดยที่มี “รูปแบบปกติ” เป็นระดับ (level) ที่อยู่ระหว่าง “สามัตถิยะภาษา” (competence) กับ “กฤตกรรมภาษา” (performance) ถ้าเรามองว่า ”สามัตถิยะภาษา” คือ คลังความรู้ (inventory) ที่เก็บรวบรวมทางเลือก (option) ทั้งหมดที่ผู้แปลมีอยู่ในหัว และ ”กฤตกรรมภาษา” คือ เซ็ตย่อย (subset) ของทางเลือกที่ผู้แปลได้เลือกใช้ “รูปแบบปกติ” ก็จะเป็นเซ็ตย่อยของ “กฤตกรรมภาษา” โดยจะเป็นตัวเลือกที่ผู้แปลมักเลือกใช้เป็นประจำ

แนวคิดในเรื่อง “รูปแบบปกติ” ทำให้เราไม่สามารถศึกษาการแปลในลักษณะที่เป็นการพิจารณางานแปลเป็นชิ้นๆ อีกต่อไปได้ แต่จำเป็นต้องศึกษาการแปลจากบทแปลจำนวนมากหรือจากคลังข้อมูลเทียบบท แนวคิดในเรื่อง “รูปแบบปกติ” นี้จึงทำให้ความสนใจของศาสตร์การแปลเปลี่ยนจากเรื่องของการคงความหมายในตัวเอกสารต้นฉบับ (equivalence) มาเป็นการศึกษาเรื่อง “รูปแบบปกติ” ของเอกสารแปล และทำให้นักวิชาการการแปลหันมาสนใจศึกษาการแปลในภาพรวมจากคลังข้อมูลเทียบบทมากกว่าที่จะศึกษาการแปลแต่ละชิ้น

5. คลังข้อมูลภาษากับศาสตร์การแปลแบบพรรณา

นับตั้งแต่ทศวรรษ 1970 ก็เริ่มมีผู้ไม่เห็นด้วยกับการศึกษาศาสตร์การแปลโดยใช้วิธีการแบบนึกคิดเอา (introspective) โดยมองว่าข้อด้อยของทฤษฎีการแปลที่มีอยู่นั้นเกิดจากการที่ไม่ได้อาศัยข้อมูลภาษาแปลที่เกิดขึ้นจริง Newman (1980) เสนอทางออกโดยให้พยายามมองหาลักษณะร่วม (generalities) ของบทแปลจากตัวอย่างจริงๆของการแปล เพื่อที่จะใช้ลักษณะร่วมดังกล่าวเป็นพื้นฐานสำหรับพัฒนาทฤษฎีหรือหาคำอธิบายอย่างเป็นระบบต่อไป และเสนอแนวคิดว่าศาสตร์การแปลควรจะเป็นการศึกษาแบบพรรณา (descriptive) มากกว่าการศึกษาแบบเดิมที่เป็นแบบสอนถูกผิด (prescriptive) แต่ผู้ที่เสนอความคิดที่เป็นรูปธรรมมากที่สุดคือ Toury โดย Toury มองว่าศาสตร์การแปลจะต้องพัฒนาสาขาของตัวเองขึ้นมาเพื่อที่วิชาด้านการแปลนี้จะได้แยกเป็นศาสตร์โดยอิสระได้ หาไม่แล้ว ศาสตร์การแปลก็จะต้องอาศัยหลักการและวิธีการจากศาสตร์อื่นๆ เช่น ภาษาศาสตร์ เพื่อสร้างกรอบการวิเคราะห์และพิสูจน์สมมติฐานอยู่ต่อไป ศาสตร์การแปลแบบพรรณา (Descriptive Translation Studies) จึงไม่ใช่เป็นเพียงการศึกษาหรือทำการวิเคราะห์เปรียบเทียบต้นฉบับและบทแปล แต่จะต้องสร้างระเบียบวิธีวิทยาวิจัยที่น่าเชื่อถือและชัดเจน เช่น Toury เสนอให้ศึกษาเงื่อนไขหรือสถานการณ์ต่างๆว่ามีผลกระทบต่อการแปล ต่อลักษณะการแปลที่เกิดขึ้นอย่างไร เพราะเหตุใดจึงเป็นเช่นนั้น เป็นต้น

การที่จะศึกษาหาลักษณะร่วมของการแปลได้จำเป็นต้องอาศัยคลังข้อมูลเทียบบท ซึ่งจริงๆแล้วก็มีผู้สนใจใช้คลังข้อมูลในการศึกษาการแปลมาก่อนหน้านี้แล้ว เช่น ในงานของ Vanderauwera (1985) ซึ่งได้จัดทำคลังข้อมูลของนวนิยาย 50 เล่มซึ่งแปลจากภาษาดัชท์เป็นภาษาอังกฤษ โดยเก็บรวบรวม

และวิเคราะห์จากคลังข้อมูลเองทั้งหมด (ไม่ได้ใช้คอมพิวเตอร์ และบอกขนาดของคลังข้อมูลเป็นจำนวนของตัวบทที่ใช้ ไม่ได้บอกเป็นจำนวนคำ) การทำงานเช่นนี้แม้จะไม่ใช่การสร้างคลังข้อมูลเทียบบทที่ใหญ่โตอะไร แต่ Vanderauwera ก็กล่าวว่าเป็นงานที่หนักและใช้เวลามาก Toury เองก็เห็นด้วยกับการใช้คลังข้อมูลเทียบบทเพื่อศึกษาหา"รูปแบบปกติ" แต่มองเห็นว่าจะมีความยุ่งยากมากขึ้นในการเลือกดึงข้อมูลและมองหาลักษณะร่วมเมื่อข้อมูลมีขนาดใหญ่มากๆ อีกทั้งยังไม่มีวิธีการทางสถิติที่ชัดเจนที่จะช่วยในการวิเคราะห์จึงจำต้องอาศัยการหยั่งรู้ (intuition) เป็นหลักในการวิเคราะห์ไปก่อน แต่ Baker (1993) แย้งว่างานทางด้านภาษาศาสตร์คลังข้อมูลที่ทำกันมาโดยเฉพาะงานของ Sinclair น่าจะให้คำตอบหลายอย่างที่น่าจะนำมาใช้ประโยชน์ได้ เพราะ Sinclair ได้จัดสร้างและใช้คลังข้อมูลภาษาขนาดใหญ่ในงานวิจัยต่างๆมาโดยตลอด และได้พัฒนาระเบียบวิธีวิจัยที่เกี่ยวข้องด้วย

Sinclair ได้กล่าวถึงความสำคัญของคลังข้อมูลภาษาต่อการแปลไว้คร่าวๆว่า คลังข้อมูลภาษาจะมามีบทบาทต่อการศึกษาเรื่องการแปลในอนาคต ซึ่งเป็นการมองว่าคลังข้อมูลเทียบบทจะเป็นประโยชน์ช่วยให้ผู้แปลสามารถแปลได้ดีมากขึ้นหรือสามารถนำมาใช้เป็นประโยชน์สำหรับระบบการแปลภาษาด้วยคอมพิวเตอร์ (machine translation) แต่ Baker (1993) คิดว่าคลังข้อมูลเทียบบทจะมีประโยชน์มากไปกว่านั้น เขาเชื่อว่า คลังข้อมูลเทียบบทไม่ได้เป็นประโยชน์เพียงแค่ทำให้รู้จักภาษานั้นๆดีขึ้น แต่ทำให้เข้าใจว่าการแปลคืออะไรและเข้าใจกลไกขั้นตอนของการแปลมากขึ้น

6. ศาสตร์การแปลและการวิเคราะห์เปรียบต่าง

ศาสตร์การแปล (translation study) ต่างจากการวิเคราะห์เปรียบต่าง (contrastive analysis) กล่าวคือ การวิเคราะห์เปรียบต่างเป็นการวิเคราะห์

เปรียบเทียบภาษาสองภาษาที่เป็นภาษาต้นฉบับทั้งคู่ แต่ศาสตร์การแปลเป็นการเปรียบเทียบภาษาสองภาษาที่เป็นภาษาต้นฉบับและเป็นภาษาแปล คลังข้อมูลสำหรับใช้ในการวิเคราะห์เปรียบต่างคือคลังข้อมูลเทียบภาษา (comparable corpora) ส่วนคลังข้อมูลสำหรับใช้ในศาสตร์การแปลคือคลังข้อมูลเทียบบท (parallel corpora) คลังข้อมูลทั้งสองแบบสามารถรวมเป็นอันเดียวได้ หากเราจัดสร้างคลังข้อมูลเทียบบทของภาษา A แปลเป็น ภาษา B และคลังข้อมูลเทียบบทของภาษา B แปลเป็นภาษา A เราก็จะได้คลังข้อมูลเทียบภาษาของภาษา A และ B ด้วย

แต่เดิมศาสตร์การแปลต้องอาศัยผลการศึกษาของการวิเคราะห์เปรียบต่างเป็นแนวทาง เพราะการวิเคราะห์เปรียบต่างเป็นการศึกษาประเด็นความเหมือนและความต่างระหว่างสองภาษา ซึ่งผลที่ได้จากการศึกษาสามารถนำมาใช้กับการแปลได้คือช่วยในการเลือกว่าจะแปลอย่างไรตามข้อจำกัดต่างๆที่มีระหว่างสองภาษานั้น แต่ปัจจุบัน มีผู้เห็นว่า การแปลเป็นเรื่องที่มีลักษณะที่ขึ้นกับบริบทและวัฒนธรรมนั้นๆด้วย จึงไม่สามารถเลือกแปลตามรูปแบบที่ได้จากการศึกษาการวิเคราะห์เปรียบต่างได้เสมอไป Schmied และ Schaffler (1996) มองความแตกต่างระหว่างการวิเคราะห์เปรียบต่างกับศาสตร์การแปลว่า การวิเคราะห์เปรียบต่างมีลักษณะที่เป็น langue-oriented ในขณะที่ศาสตร์การแปลมีลักษณะที่เป็น parole-oriented และในขณะที่ผลที่ได้จากการวิเคราะห์เปรียบต่างเป็นเรื่องของโครงสร้างที่ดูเป็นธรรมชาติในภาษานั้นๆ ผลที่ได้จากศาสตร์การแปลจะทำให้เห็นลักษณะเฉพาะที่ปรากฏในบทแปลเท่านั้น ซึ่งเบี่ยงเบนไปจากภาษาปกติโดยอิทธิพลของภาษาต้นฉบับ Schmied และ Schaffler (1996) เรียกลักษณะเฉพาะของเอกสารแปลนี้ว่า translationese

7. การศึกษาลักษณะเฉพาะของภาษาแปล

Schmied และ Schaffler (1996) มองว่า ลักษณะเฉพาะของภาษาแปลหรือ translationese มีสองแบบ แบบแรก คือการเบี่ยงเบนจากระบบ (system) และการเบี่ยงเบนจากรูปแบบปกติ (norm) ของภาษานั้น ซึ่งเป็นเพราะอิทธิพลของภาษาต้นฉบับ ส่วนลักษณะเฉพาะของเอกสารแปลแบบที่สอง คือลักษณะที่จัดได้ว่าเป็นสากลลักษณ์ของภาษาแปล (universal feature) เองไม่ได้เป็นผลมาจากอิทธิพลของภาษาต้นฉบับ

7.1 การศึกษาการเบี่ยงเบนจากระบบและรูปแบบปกติ

ลักษณะที่พบในภาษาแปลที่เป็นการเบี่ยงเบนจากระบบหรือบางครั้งเรียกว่าเป็น system error คือมีการเลือกใช้ลักษณ์ (feature) ที่ถือได้ว่าเป็นการใช้ที่ผิดหลักไวยากรณ์ในภาษานั้น ซึ่งเป็นอิทธิพลที่ได้รับมาจากภาษาต้นฉบับ ส่วนการเบี่ยงเบนจากรูปแบบปกติ คือ ลักษณะที่เป็นการเบี่ยงเบนการใช้จากรูปแบบปกติในภาษานั้น ซึ่งก็เป็นอิทธิพลที่ได้มาจากภาษาต้นฉบับเช่นกัน ตัวอย่างเช่น ในการศึกษาภาษาเยอรมันกับภาษาอังกฤษของ Schmied และ Schaffler ภาษาเยอรมันซึ่งมีแนวโน้มที่จะพูดแบบตรงไปตรงมามากกว่าภาษาอังกฤษ เมื่อศึกษาดูลักษณ์ทางภาษาที่เกี่ยวข้องกับเรื่อง tentativeness หรือ indirectness เช่น ดู semi-modal construction (เช่น *seem to, tend to*) ดู modal adverb และ particle (เช่น *perhaps*, *possibly*, *eher*, *wohl*) เป็นต้น ก็จะพบว่าในภาษาอังกฤษปกติจะมีการใช้ลักษณ์เหล่านี้มากกว่าในภาษาเยอรมัน และเมื่อเปรียบเทียบตัวบทภาษาเยอรมันที่ได้จากการแปลจากภาษาอังกฤษ ก็พบว่าถึงแม้จะมีการใช้ลักษณ์ที่บอกความเป็น tentativeness หรือ indirectness น้อยกว่าภาษาอังกฤษ แต่ก็มีการใช้ลักษณ์เหล่านี้มากกว่าในภาษาเยอรมันปกติ จึงถือได้ว่าเป็นผลมาจากอิทธิพลของภาษาต้นฉบับคือภาษาอังกฤษ

7.2 การศึกษาสากลลักษณ์ของการแปล

Baker (1993) กล่าวว่างานสำคัญที่สุดที่รอคอยการนำคลังข้อมูลภาษามาใช้ในศาสตร์การแปลคือการค้นหาธรรมชาติของการแปลในฐานะที่เป็นกระบวนการสื่อสาร (communicative event) ชนิดหนึ่ง สิ่งหนึ่งที่สามารถกระทำได้คือการศึกษาหาสากลลักษณ์ (universal features) ของการแปล หรือคือการหาลักษณ์ทางภาษาที่จะปรากฏเฉพาะในบทแปลไม่ว่าจะเป็นการแปลในภาษาใด ตัวอย่างของลักษณ์เฉพาะของการแปลที่ Baker ยกมาเป็นตัวอย่าง เช่น

- การให้ข้อมูลเพิ่มเติมเพื่อความชัดเจน เช่น ในบทแปลจะให้รายละเอียดข้อมูลเสริม (background information) ที่ไม่ปรากฏในตัวต้นฉบับ เนื่องจากผู้แปลอาจจะเกรงว่าผู้อ่านไม่มีพื้นความรู้พอที่จะเข้าใจในเรื่องนั้น เช่น ประโยคว่า *The example of Truman was always present in my mind* มีผู้แปลเป็นภาษาอารบิคในความหมายต่อไปนี้ in my mind there was always the example of the American President Harry Truman, who succeeded Franklin Roosevelt towards the end of World War II. At that time – and after Roosevelt – truman seemed a rather nondescript and unknown character who could not lead the great human struggle in World War II to its desired and inevitable and. But Truman – faced with the challenge of practical experience – grew and matured and became one of the most prominent American presidents in modern times. I imagined that the same thing could happen to Sadat.
- แนวโน้มที่จะขจัดความกำกวมและทำให้ง่ายขึ้น คือ การทำให้ความกำกวมที่แฝงอยู่ในต้นฉบับหายไป ไม่ปรากฏในบทแปล เช่น ในคลังข้อมูลภาษาอังกฤษที่แปลมาจากภาษาดัชท์ซึ่ง Vanderauwera รวบรวมจากนว

นิยายนั้น คำสรรพนามที่กำกวมในภาษาดัชท์จะถูกแทนด้วยคำที่ไม่กำกวม และรูปประโยคที่ยากก็จะถูกทำให้ง่ายขึ้น

- การทำให้ถูกไวยากรณ์ คือ การทำให้ประโยคที่ผิดไวยากรณ์ในต้นฉบับให้เป็นประโยคที่ถูกไวยากรณ์ในบทแปล
- การหลีกเลี่ยงการซ้ำคำ ซ้ำประโยค (repetition) ที่ปรากฏในต้นฉบับ
- การเพิ่มหรือขยายลักษณ์บางอย่างในภาษาแปล คือ ภาษาในบทแปลมีลักษณ์บางอย่างที่ปกติจะไม่พบในภาษาปกติ ตัวอย่างเช่น Shamaa (1978) รายงานไว้ว่า ในภาษาอังกฤษที่แปลมาจากภาษาอารบิคจะมีคำว่า *say* และ *day* มากกว่าที่จะพบในภาษาอังกฤษปกติ แต่ถึงกระนั้น คำสองคำนี้ก็นับว่า ปรากฏน้อยเมื่อเทียบกับคำที่มีความหมายเดียวกันในภาษาอารบิค ปรากฏการณ์นี้เป็นที่รู้จักในชื่อของ “the third code” ซึ่งเป็นผลจากการเผชิญหน้ากันระหว่างภาษาของเอกสารต้นฉบับและภาษาของเอกสารแปล และเป็นเครื่องชี้ให้เห็นความแตกต่างระหว่างภาษาตามปกติทั้งสองกับภาษาแปล (ปรากฏการณ์นี้ก็คือการเบี่ยงเบนจากรูปแบบปกตินั่นเอง)

ลักษณะต่างๆที่กล่าวมานี้เป็นลักษณะที่ทำให้เอกสารแปลต่างจากเอกสารปกติในภาษานั้นๆ แต่อย่างไรก็ดี Baker ยอมรับว่าคงต้องใช้เวลาบ้างเพื่อให้มีวิธีวิทยาวิจัยทางคลังข้อมูลภาษา (corpus methodology) ที่ดีพอที่จะใช้ศึกษาสากลลักษณ์เหล่านี้ แต่ในขั้นนี้ Baker ก็เสนอว่าเราน่าจะมองการศึกษาตรงนี้ออกได้ในระดับหนึ่ง เช่น หากว่าต้องการศึกษาการแปลจากภาษาต่างๆ มาเป็นอังกฤษ เราอาจใช้คลังข้อมูลภาษาหารูปแบบต่อไปนี้

- เปรียบภาษาอังกฤษที่ได้จากการแปลแบบต่างๆ หารูปแบบของภาษาอังกฤษที่แปลมาจากภาษาอื่นๆว่ามีลักษณะร่วมกันอะไรบ้าง
- เปรียบภาษาอังกฤษที่ได้จากการแปลกับภาษาอังกฤษธรรมดาเพื่อศึกษาว่ามี

รูปแบบอะไรที่ภาษาอังกฤษที่ได้จากการแปลมีหรือไม่มี

นอกจากการใช้คลังข้อมูลเทียบบทเพื่อศึกษาสากลลักษณ์ของการแปลแล้ว เรายังสามารถใช้คลังข้อมูลเทียบบทเพื่อศึกษาลักษณ์ที่ไม่ได้มีลักษณะเป็นสากล (universal) แต่เป็นลักษณ์ที่แตกต่างกันตามแต่ละวัฒนธรรมและสังคม ตัวอย่างเช่น ภาษาฮิบรูจะไม่นิยมใช้ชื่อหัวเรื่อง (title) ที่ซับซ้อนไป (Toury 1980) ในภาษาดัชท์นิยมแทนคำต่างประเทศด้วยคำของภาษาดัชท์เอง (Vanderauwera 1985) ในภาษาญี่ปุ่นยอมให้มีการใช้คำยืม (loan word) ได้ง่าย (Baker 1992) เป็นต้น

Baker กล่าวว่า ในขณะนี้ ในสหราชอาณาจักรประเทศอังกฤษ มีการเสนอให้ผู้แปลได้ให้ข้อมูลต่างๆเกี่ยวกับการแปลตามรูปแบบ (format) ที่กำหนดเวลาที่มีการแปลเอกสารใดใด ถ้าทำได้ตามนี้ ก็จะได้คลังข้อมูลเทียบบทที่มีข้อมูลต่างๆที่จะใช้ศึกษาเรื่องต่างๆ เช่น ความต่างทางเพศ อายุ ของผู้แปล เป็นต้น กล่าวโดยสรุปก็คือ ศาสตร์การแปลกำลังอยู่ในช่วงของการพัฒนาศาสตร์ของตัวเองขึ้นมา และคลังข้อมูลภาษาเป็นปัจจัยจำเป็นอันหนึ่งสำหรับพัฒนาการของศาสตร์ด้านนี้ ซึ่งเรียกว่า การศึกษาการแปลโดยใช้คลังข้อมูลภาษา (Corpus–Based Translation Studies)

8. การใช้คลังข้อมูลภาษาเพื่อช่วยในการแปล

โดยทั่วไปแล้ว งานแปลที่จะประสบความสำเร็จได้ดีนั้นผู้แปลควรจะต้องมีเงื่อนไขความพร้อม 3 ประการ คือ ประการแรก ผู้แปลจะต้องมีความรู้ในภาษาต้นฉบับเป็นอย่างดี ประการที่สอง ผู้แปลต้องมีความเข้าใจในเนื้อหาของเอกสารที่แปลเป็นอย่างดี เพราะความรู้ทางภาษาอย่างเดียวไม่ได้เป็นเครื่องยืนยันว่าจะทำให้ผู้แปลอ่านและเข้าใจเอกสารได้ถูกต้อง ประการที่สามคือผู้แปลต้องมีความสามารถในการใช้ภาษาสำหรับแปลได้ดี ซึ่งก็เป็นที่เชื่อกันว่าการแปล

ที่ดีควรเป็นการแปลจากภาษาที่สองมาเป็นภาษาแม่ เนื่องจากผู้แปลจะสามารถแปลได้ดีกว่า แต่ถึงแม้การแปลที่เกิดขึ้นส่วนมากจะเป็นการแปลจากภาษาที่สองมาเป็นภาษาแม่ โอกาสที่ผู้แปลจะต้องแปลจากภาษาแม่มาเป็นภาษาที่สองก็อาจเกิดขึ้นได้ ในกรณีเช่นนี้ ผู้แปลอาจประสบปัญหาในเรื่องของการเลือกใช้คำหรือสำนวนในภาษาที่สองให้เป็นไปตามธรรมชาติที่ผู้พูดภาษานั้นใช้กัน หรือแม้แต่ในกรณีที่ผู้แปลแปลจากภาษาที่สองมาเป็นภาษาแม่ ผู้แปลก็อาจจะต้องแปลเอกสารเฉพาะทางในสาขาที่ผู้แปลไม่มีความรู้ดีพอ ซึ่งในกรณีเช่นนี้ นอกจากปัญหาในเรื่องการทำความเข้าใจกับเนื้อหาแล้ว ผู้แปลก็ยังอาจมีปัญหาในการเลือกใช้ศัพท์หรือสำนวนให้สอดคล้องกับรูปแบบภาษาที่ผู้เชี่ยวชาญในสาขานั้นใช้กัน ทำให้ไม่สามารถแปลเอกสารได้ใกล้เคียงภาษามาตรฐานที่ใช้กันในสาขานั้น การใช้คลังข้อมูลภาษาสามารถช่วยแก้ปัญหาเหล่านี้ได้ ซึ่งประโยชน์ของคลังข้อมูลภาษาในการแปลสามารถแยกได้เป็นสองลักษณะ คือ ประโยชน์ต่อการสอนแปล และประโยชน์ในลักษณะที่เป็นทรัพยากรช่วยในการแปล

8.1 การใช้คลังข้อมูลในการสอนแปล

Bowker (1998) กล่าวจากประสบการณ์ของการสอนแปลที่ผ่านมาว่าปัญหาที่พบในการแปลนั้น ข้อผิดพลาดที่เกิดขึ้นมีทั้งที่เกิดมาจากการทำความเข้าใจ (comprehension) และที่เกิดจากการเขียนคำแปล (production) ข้อผิดพลาดจากความเข้าใจเกิดจากการขาดความรู้ในสาขานั้นมากกว่าที่จะเกิดจากการขาดความรู้ในภาษาต้นฉบับ ส่วนข้อผิดพลาดจากการเขียนคำแปลเกิดจากการขาดความรู้ (competence) ในภาษาเฉพาะสาขานั้น หรือการขาดความรู้ในการใช้ภาษาแปล ทำให้เลือกใช้ศัพท์ผิด หรือเลือกใช้สำนวนหรือลีลาที่ไม่เหมาะสม การใช้คำศัพท์ผิดนั้นยังพอแก้ไขได้ง่ายกว่า เพราะเมื่อแปลงานในสาขานั้นไปบ่อยๆ ผู้แปลก็จะรู้ว่าควรใช้ศัพท์อะไร หรืออาจใช้พจนานุกรมเฉพาะ

สาขาช่วยก็ได้ แต่การเลือกใช้สำนวนหรือเลือกใช้คำต่างๆร่วมกัน (collocation) ผิดนั้นแก้ไขได้ยากกว่า เพราะไม่มีทรัพยากรโดยตรงที่ผู้แปลจะใช้อ้างอิงได้ เช่น ผู้แปลอาจแปลคำว่า *ชาแก่* ว่า *old tea* หรือ *powerful tea* แทนที่จะแปลว่า *strong tea* หากผู้แปลจะแปลให้ตรงตามรูปแบบที่ใช้กันจริง ผู้แปลก็จะต้องเสียเวลามากในการหาเอกสารต้นฉบับที่เขียนในภาษาในสาขานั้นๆมาอ่านเพื่อเลือกใช้คำแปลให้เหมาะสม ไม่เช่นนั้น ผู้แปลก็ต้องยอมให้งานแปลออกมาแปร่งจากมาตรฐาน ปัญหานี้จะพบทั้งในการแปลจากภาษาแม่เป็นภาษาที่สอง (L1->L2) หรือจากภาษาที่สองเป็นภาษาแม่ (L2->L1) ในการแปลแบบ L1->L2 นั้นแน่นอนว่าปัญหาเกิดจากการที่ผู้แปลไม่มีความรู้ในภาษา L2 ดีพอที่จะเลือกศัพท์หรือคำที่เกิดร่วมกัน (collocation) ที่เหมาะสมเหมือนผู้พูดภาษานั้นได้ ส่วนในการแปล L2->L1 นั้นแม้จะไม่มีปัญหาเพราะเป็นภาษาแม่ของผู้แปลเอง แต่งานแปลที่ได้รับ อาจเป็นงานเขียนเฉพาะสาขา เช่น เอกสารแพทย์ เอกสารทางเทคนิค ซึ่งการใช้ภาษาให้คล้องตามที่ผู้เชี่ยวชาญในสาขานั้นใช้กันก็เป็นเรื่องยาก เพราะภาษาเฉพาะในสาขากลายเป็นเสมือนภาษาที่ผู้แปลไม่คุ้นเคยดี ต้องรู้ว่าจะใช้ศัพท์ สำนวนให้สอดคล้องกับที่ผู้เชี่ยวชาญในสาขาใช้กันอยู่ได้อย่างไร การพยายามสอนนักแปลเพื่อแก้ไขจุดอ่อนนี้ โดยการเพิ่มรายวิชาที่เกี่ยวข้องกับภาษาแม่หรือภาษาที่สองนั้นทำได้ยากด้วยข้อจำกัดของเวลาในหลักสูตร ส่วนการสอนการแปลโดยเน้นให้มีประสบการณ์การแปลกับทุกสาขานั้นก็เป็นไปได้ยาก เพราะไม่สามารถคาดเดาได้ว่านักแปลจะต้องแปลเอกสารประเภทไหนบ้าง สอนได้แต่เพียงหลักการกว้างๆ หรือการพยายามสอนให้นักแปลมีความรอบรู้ในเนื้อหาสาขาทุกๆด้าน ก็เป็นไปไม่ได้เช่นกัน เพราะเวลาของหลักสูตรไม่ได้มีมากพอ ทางออกคือการสร้างทรัพยากรและสอนให้นักแปลรู้จักวิธีในการหาความรู้ทางภาษาและเนื้อหาเพิ่มเติมเอาเอง ซึ่งหนทางหนึ่งที่สามารถทำได้คือการสอนให้นัก

แปลรู้จักใช้ประโยชน์จากคลังข้อมูลภาษา

นอกจากนี้ Bernardini (1997) มองว่าการแปลที่ดีนั้นจำเป็นต้องอาศัยการพัฒนาศักยภาพเฉพาะที่เกี่ยวข้องกับการแปล 3 ด้าน ได้แก่ awareness, reflectiveness และ resourcefulness

1. awareness คือศักยภาพในการตระหนักว่าความหมายมีมากกว่าที่เห็นจากตัวอักษร ผู้แปลควรจะสามารถมองทะลุตัวบท เข้าใจว่าทำไมผู้เขียนจึงเขียนออกมาในลักษณะนั้น ผู้แปลควรจะสามารถมองเห็นภาษาเป็นภาพเครือข่ายของการเลือกใช้ภาษาแบบต่างๆที่สัมพันธ์กัน มีอิทธิพลและได้รับอิทธิพลจากวัฒนธรรมสังคมนั้น เข้าใจกฎต่างๆที่ครอบคลุมการใช้ เข้าใจความหมายเพิ่มเติมอื่นๆเมื่อมีการละเมิดกฎเหล่านั้น เข้าใจว่ามีทั้งสิ่งที่เป็นความสัมพันธ์ระหว่างข้อความ (intertextual) และความสัมพันธ์ภายในข้อความ (intratextual) และเข้าใจกลวิธีต่างๆที่จะใช้ในการแปล

2. reflectiveness คือศักยภาพในการวิเคราะห์ตัวบท สามารถตีความความหมายแฝง (implication) ต่างๆ เช่น เจตนาและความตั้งใจของผู้พูด เห็นเหตุผลของการใช้กลวิธีต่างๆในการเขียน เช่น ทำไมจึงใช้ตัวเอียง ตัวหนา เป็นต้น

3. resourcefulness คือศักยภาพที่จะแสวงหาความรู้ใหม่ที่จำเป็นเพิ่มเติม (knowing how to find what where)

Bernardini เสนอว่าวิธีการสอนแปลนั้นต้องตั้งจุดมุ่งหมายเพื่อพัฒนาศักยภาพทั้งสามอย่างนี้ และว่าการใช้คลังข้อมูลภาษาสำหรับการเรียนจะช่วยพัฒนาศักยภาพทั้งสามนี้ได้ โดย Bernardini กล่าวถึงวิธีการสอนแปลที่ผ่านมาคือการสอนแปลแบบดั้งเดิม (traditional) ว่าเป็นการให้แบบฝึกหัดผู้เรียนไปแปลโดยมีอาจารย์ผู้สอนตรวจแก้ไขและให้คำวิจารณ์ (feedback) ซึ่งอาจไม่ใช่วิธีการสอนแบบที่มีประสิทธิภาพมากที่สุด การสอนแบบนี้เป็นการผลักภาระการพัฒนา

ทักษะ (skill) ไปให้กับผู้เรียน การตรวจแก้ไขกลับมาให้ก็ไม่ได้ทำให้มีการพัฒนา awareness หรือทำให้ผู้เรียนได้กลวิธีต่างๆ (export strategy) ของการแปลได้ อีกทั้ง พื้นฐานแนวการเรียนการสอน (pedagogic foundation) ของการสอนแปลแบบนี้จะค่อนข้างอ่อน เพราะเป็นเพียงการให้โอกาสได้ฝึกหัดแปล และหวังว่าผู้เรียนจะสามารถพัฒนาฝีมือการแปลได้เองแล้ว (หากเป็นเช่นนั้นจริง ผู้เรียนก็อาจจะไม่ต้องมาเรียนในหลักสูตรการแปล อาศัยการทำฝึกหัดการแปลอย่างสม่ำเสมอ และหาคนช่วยตรวจแก้ให้ ก็น่าจะได้ผลดีเช่นกัน)

Bernardini มองว่า แม้เรื่องการแปลเป็นเรื่องที่เกี่ยวกับการใช้ภาษาเฉพาะทาง (LSP) แต่ก็ไม่เหมือนกับ LSP เสียทีเดียว จึงไม่สามารถใช้วิธีการเรียนการสอนแบบที่ใช้กันใน LSP ได้ โดย Bernardini อ้างถึงความแตกต่างระหว่างการฝึกหัด (training) กับการศึกษา (education) ตามที่ Widdowson (1983,1984) ได้แบ่งไว้ โดยที่การฝึกหัดเป็นการคาดเดาว่าจะพบกับปัญหาอะไรแล้วฝึกให้แก้ปัญหานั้น LSP มีลักษณะที่เป็นการฝึกหัด หลักสูตรการสอนจะเน้นที่การเรียนรู้ความรู้ต่างๆที่จะใช้กับสถานการณ์ที่กำหนด วัตถุประสงค์และเป้าหมายจะพ้องกัน (วัตถุประสงค์ เป็นสิ่งที่บรรลุหลังจากจบวิชา เป้าหมายเป็นสิ่งที่บรรลุในระยะยาว) เช่น เพื่อให้ใช้ภาษาสำหรับงานวิศวกรรมยานยนต์ได้ การเรียนแบบนี้มีลักษณะเป็นกระบวนการสะสม (cumulative process) ส่วนการศึกษาจะเน้นที่การพัฒนาศักยภาพทางปัญญาทั่วไป (general intellectual capacity) และความสามารถต่างๆเพื่อจะสามารถแก้ปัญหาที่อาจจะพบในอนาคต การเรียนภาษาเพื่องานทั่วๆไป (general purpose language learning) จะมีลักษณะที่เป็นการศึกษา หลักสูตรการสอนจะเน้นที่การพัฒนาศักยภาพ (capacity) หรือความสามารถในการแสวงหาองค์ความรู้ที่ต้องการ วัตถุประสงค์และเป้าหมายไม่จำเป็นต้องเป็นอย่างเดียวกัน วัตถุประสงค์ของ

หลักสูตรอาจจะนำไปสู่เป้าหมายที่ต้องการ เช่น วัตถุประสงค์ในหลักสูตรคือให้สามารถแปลเอกสารที่กำหนดให้ เป้าหมายคือให้สามารถแปลเอกสารอะไรก็ได้ การเรียนแบบนี้มีลักษณะที่เป็นกระบวนการสร้างสรรค์ (generative process) Bernardini มองว่าการสอนแปลมีจุดมุ่งหมายแบบเดียวกับการศึกษามากกว่าการฝึกหัด และเป็นไปเพื่อพัฒนาศักยภาพบางอย่าง

การฝึกหัด (LSP)	การศึกษา (LGP)
- เป้าหมาย (aims) ตรงกับวัตถุประสงค์ (objectives)	- เป้าหมาย (aims) ไม่ตรงกับวัตถุประสงค์ (objectives) (วัตถุประสงค์นำไปสู่เป้าหมาย)
- เพื่อการเพิ่มพูน	- เพื่อการพัฒนา
- เน้นที่ความรู้ (competence) (ความรู้ที่จะนำมาใช้)	- เน้นที่ศักยภาพ (ความสามารถที่จะแสวงหาความรู้)
- เป็นการสั่งสม (cumulative)	- เป็นการสร้างสรรค์ (generative)
- ผลงานเป็นสิ่งสำคัญ (product-oriented)	- กระบวนการเป็นสิ่งสำคัญ (process-oriented)

ตาราง 4.1 เปรียบเทียบระหว่างการฝึกหัดกับการศึกษา (Bernardini 1997)

ในประเด็นเรื่อง LSP นี้ Gavioli (1996) ก็มองว่าการแปลเป็นการใช้ภาษาเฉพาะทาง แต่ก็แตกต่างจากการใช้ภาษาเฉพาะทางแบบปกติ (non-standard LSP) เพราะผู้เรียนแปลเป็นผู้เชี่ยวชาญด้านภาษามากกว่าจะเป็นผู้เชี่ยวชาญสาขาเฉพาะ แต่ก็ต้องการความรู้เฉพาะทาง (technical competence) ซึ่งการที่จะได้ความรู้นี้มาก็ไม่อยู่ในวัตถุประสงค์ของหลักสูตร แต่เนื่องจากการที่ไม่สามารถคาดได้ว่าจะต้องแปลเอกสารสาขาใด การเรียนการ

สอนจึงไม่ใช่ LSP แบบทั่วไป ซึ่งสอดคล้องกับข้อสรุปของ Bernardini (1997) ที่ว่า LSP เป็นเรื่องการเรียนรู้ความรู้เฉพาะทาง (acquire specialized competence) ส่วนการแปลต้องการการพัฒนาศักยภาพเฉพาะที่เกี่ยวข้องกับการแปล อันประกอบด้วยศักยภาพ 3 ด้าน คือ awareness reflectiveness และ resourcefulness

Bernardini มองว่าวิธีการสอนแปลโดยใช้คลังข้อมูลภาษานี้ไม่ใช่การสอนแบบให้ฝึกหัดทำงาน (task-based) แบบการสอนที่เป็นมา เพราะไม่มีการกำหนดบท (script) ไว้ล่วงหน้าว่าผู้เรียนจะพบอะไร ขึ้นอยู่กับผู้เรียนเองเป็นสำคัญ จึงไม่มีการให้เกรดงานที่ทำ และมีลักษณะที่เป็นจริง (authentic) มากกว่าเพราะไม่ได้เป็นการเลียนแบบ (replicate) หรือจำลอง (simulate) เหตุการณ์อะไร Bernardini มองว่าการเรียนแบบใหม่นี้กระตุ้นและท้าทายให้ผู้เรียนอยากเรียน และสนุกกับการเรียน ผู้เรียนจะใช้เวลากับการสำรวจข้อมูลในคลังข้อมูลนานเท่าใดก็ได้ ผู้เรียนมีอิสระที่จะตั้งคำถามเอง เลือกวิธีการที่จะให้ได้มาซึ่งคำตอบ เช่น จะค้นจากคำอะไร จะดูผลที่ได้จากโปรแกรมคอนคอร์แด๊นซ์อย่างไร จะทำงานคนเดียวหรือทำงานเป็นกลุ่ม จะเขียนสรุปผลอย่างไร งานแบบนี้มีลักษณะที่เป็นการแสวงหาค้นคว้า (journeys of discovery) เพราะคาดเดาไม่ได้ว่าจะได้ผลออกมาอย่างไร อาศัยความสามารถของผู้เรียนในการทำความสำคัญไม่ได้อยู่ที่จุดเริ่มและจุดสิ้นสุด แต่อยู่ที่สิ่งที่ได้พบระหว่างทาง การเรียนจากคลังข้อมูลภาษาแบบนี้ คือสิ่งที่ John (1986) เรียกว่าการเรียนรู้จากข้อมูล (data-driven learning) ดังนั้น Bernardini จึงมองการสอนแปลในลักษณะนี้ว่าเป็นการเรียนรู้เพื่อที่รู้กระบวนการ (knowledge how to) มากกว่าที่จะรู้ความรู้เรื่องใดเรื่องหนึ่ง (knowledge that) จุดสนใจอยู่ที่กระบวนการ (process) ไม่ใช่ผลที่จะได้ (product) การเรียนการสอนแบบนี้จะช่วยให้เกิดการ

พัฒนาศักยภาพทั้งสามอย่าง คือ awareness reflectiveness และ resourcefulness Bernardini ยกตัวอย่างของคำว่า *good old* ว่าดูเสมือนเป็นคำที่เข้าใจง่าย ไม่น่ามีปัญหาอะไร แต่หากตั้งคำถามเกี่ยวกับ *good old* ในลักษณะต่อไปนี้ เช่น คำอะไรที่มักจะปรากฏร่วมกับ *good old* การใช้สำนวนนี้มีความหมายแฝง (implication) อะไร เวลาใช้อยู่ในบริบทเป็นอย่างไร เช่น มีรูปแบบโครงสร้างประโยคแบบไหน มีการใช้เครื่องหมายวรรคตอนประกอบไหม *good old* ใช้ในทางตรงข้ามกับ *bad old* หรือไม่ คำถามเหล่านี้จะตอบได้อย่างไร การจะตอบคำถามเหล่านี้ได้จำเป็นต้องอาศัยคลังข้อมูลภาษา

กล่าวโดยสรุป Bernardini กล่าวว่าการใช้คลังข้อมูลภาษาได้ช่วยพัฒนาศักยภาพทั้งสามด้าน ดังนี้ awareness การใช้คลังข้อมูลภาษาช่วยให้เกิดความสงสัยต่อสิ่งที่เห็นชัดเจน คำที่คิดว่าเข้าใจดีก็เกิดความสงสัย เกิดความตระหนักว่าอาจมีบางอย่างที่ยังไม่รู้ไม่เข้าใจได้

reflectivness การใช้คลังข้อมูลภาษาช่วยให้มีความรู้ความเข้าใจในการใช้ภาษานั้นๆดีขึ้น ทำให้รู้เรื่องเกี่ยวกับคำต่างๆมากขึ้น (lexical competence) ที่สำคัญมีความชำนาญที่จะแสวงหาความเข้าใจได้เองต่อไป

resourcefulness แน่นอนว่า การใช้คลังข้อมูลภาษาทำให้ผู้เรียนสามารถค้นคว้าหาความรู้ต่างๆเพิ่มเติมเองในภายหลังได้

8.2 การใช้คลังข้อมูลเป็นทรัพยากรการแปล

ทรัพยากรทั่วไปที่ใช้ในการแปลนั้น ได้แก่ พจนานุกรมและสารานุกรมต่างๆ พจนานุกรมที่ใช้เป็นได้ทั้ง พจนานุกรมทั่วไปและพจนานุกรมเฉพาะทาง ข้อเสียของพจนานุกรมทั่วๆไปคือมโนทัศน์ (concept) ต่างๆ ถูกจัดแยกจากกัน ส่วนพจนานุกรมเฉพาะทาง แม้บางเล่มอาจให้ข้อมูลคำที่ใช้ร่วมกัน (collocation) บ้างแต่ยังไม่เพียงพอ การจะเข้าใจความหมายของคำได้ดี ก็เหมือนกับการที่จะ

เข้าใจพฤติกรรมของสัตว์ คือเราต้องเฝ้าดูหรือศึกษาในธรรมชาติแวดล้อมที่สัตว์นั้นอาศัยอยู่ ดังนั้น การจะเข้าใจพฤติกรรมการเกิดของคำ ก็ต้องดูจากสิ่งแวดล้อมธรรมชาติหรือคือบริบทต่างๆที่คำนั้นปรากฏ การใช้คลังข้อมูลภาษาจะช่วยให้เห็นสิ่งเหล่านี้ ซึ่งเป็นสิ่งที่พจนานุกรมต่างๆไม่มีให้

Friedbichler (1997) เสนอให้ใช้ทรัพยากรผสมระหว่างพจนานุกรม คลังข้อมูลภาษา และอินเทอร์เน็ต (Lexicon-Corpus-Web resources) สำหรับใช้ในการแปล ทรัพยากรแรกคือพจนานุกรมซึ่งเหมาะสำหรับใช้กับศัพท์พื้นฐานต่างๆ (basic terminology) ถ้าทรัพยากรนี้ไม่เพียงพอจึงใช้คลังข้อมูลภาษาช่วย Friedbichler เสนอให้ใช้คลังข้อมูลภาษาขนาด 500,000 – 5 ล้านคำโดยอ้างว่าจะครอบคลุมถึง 97% ของคำในภาษา ที่เหลือจากนั้น คำที่แปลกมากๆจึงใช้อินเทอร์เน็ตเป็นทรัพยากรช่วยในการแปล Friedbichler เชื่อว่าคลังข้อมูลภาษาเป็นประโยชน์อย่างมากต่อการแปลภาษาเฉพาะทาง และสถาบันที่สอนการแปลต้องสอนให้นักแปลใช้ทรัพยากรเหล่านี้เป็น

แต่ข้อเสียของการใช้คลังข้อมูลภาษาในงานทั่วๆไป คือการสร้างคลังข้อมูลภาษาขึ้นมานั้นต้องใช้เวลาโดยเฉพาะถ้าต้องการคลังข้อมูลภาษาที่มีการออกแบบโครงสร้างอย่างดี แต่สำหรับนักแปลแล้ว การที่นักแปลจะรวบรวมคลังข้อมูลภาษาเพื่อวัตถุประสงค์ของนักแปลเองทำได้ง่ายกว่า เพราะสามารถหาบทความในภาษาเป้าหมายที่เขียนโดยผู้เชี่ยวชาญในสาขานั้น นำเอาบทความที่ถูกอ้างถึงในเอกสารที่จะแปล ค่อยๆเก็บสะสมไว้ ในระยะยาวก็จะได้คลังข้อมูลที่ต้องการ อีกทั้งมีเอกสารจำนวนมากที่อยู่ในอินเทอร์เน็ต อยู่ในซีดีรอม แต่การเลือกเอกสารจากอินเทอร์เน็ตก็ต้องระวังเพราะเอกสารบางชิ้นก็อาจจะไม่น่าเชื่อถือเนื่องจากใครก็สามารถเขียนข้อความต่างๆเข้าไว้ในอินเทอร์เน็ตได้

คลังข้อมูลภาษาที่จะเป็นประโยชน์สำหรับนักแปลเป็นได้ทั้งคลังข้อมูลภาษาเดียว (monolingual corpora) หรือคลังข้อมูลสองภาษา (bilingual corpora) และเป็นได้ทั้งคลังข้อมูลภาษาทั่วไป (general corpora) หรือคลังข้อมูลภาษาเฉพาะด้าน (specialized corpora) ในส่วนของคลังข้อมูลสองภาษานั้นแบ่งได้เป็นคลังข้อมูลเทียบบท (parallel corpora) และคลังข้อมูลเทียบภาษา (comparable corpora) คลังข้อมูลเทียบบทเป็นคลังข้อมูลของทั้งภาษาต้นฉบับและภาษาแปล ซึ่งมีการจับคู่ข้อความ (alignment) ระหว่างภาษาต้นฉบับและภาษาแปล เช่น จับคู่ระหว่างประโยคของภาษาต้นฉบับกับประโยคภาษาแปล ทำให้รู้ว่าประโยคแบบนี้แปลออกมาเป็นประโยคในอีกภาษาหนึ่งอย่างไร ส่วนคลังข้อมูลเทียบภาษาเป็นคลังข้อมูลของสองภาษา โดยข้อมูลภาษาทั้งสองนั้นเป็นภาษาต้นฉบับทั้งคู่ แต่สามารถเปรียบเทียบกันได้ เพราะใช้เกณฑ์ในการคัดเลือกเอกสารแบบเดียวกัน เช่น คลังข้อมูล ICE (International Corpus of English) เป็นคลังข้อมูลเทียบภาษา เพราะคลังข้อมูลแต่ละอันเป็นภาษาอังกฤษในแต่ละประเทศ และเป็นคลังข้อมูลที่มีโครงสร้างแบบเดียวกันคือเก็บเอกสารชนิดเดียวกันในสัดส่วนที่เหมือนกัน

คลังข้อมูลการแปลเป็นประโยชน์อย่างมากต่อศาสตร์การแปลดังที่กล่าวมาแล้ว คือทำให้เราสามารถศึกษาการแปลเพื่อที่จะเข้าใจธรรมชาติของการแปล มองเห็นรูปแบบต่างๆ และรูปแบบปกติของการแปล หลายๆ คน มองว่าคลังข้อมูลการแปลมีประโยชน์ต่อการแปล ทำให้เห็นตัวอย่างการแปล สามารถค้นดูการแปลที่เคยใช้กัน (โปรแกรมช่วยแปลในปัจจุบัน เช่น Star Transit, Trados ต่างก็ใช้ประโยชน์จากคลังข้อมูลเทียบบทในการช่วยแปล โดยจัดเก็บคลังข้อมูลเทียบบทที่ได้จากแปลครั้งก่อนๆเป็นส่วนของฐานข้อมูลที่เรียกว่า Translation Memory) แต่คลังข้อมูลภาษาเดียวและคลังข้อมูลเทียบภาษาก็เป็นประโยชน์ต่อ

นักแปลเช่นกัน เพราะสามารถใช้เป็นทรัพยากรการแปลอย่างหนึ่ง Friedbichler (1997) เห็นว่าการใช้คลังข้อมูลเทียบภาษาเป็นประโยชน์มากกว่าการใช้คลังข้อมูลเทียบบท โดยมองว่า ในคลังข้อมูลเทียบบทมีความแตกต่างพื้นฐานทางภาษาศาสตร์ระหว่างภาษาต้นฉบับและภาษาแปลอยู่ ตัวเอกสารต้นฉบับก็ใช่ว่าจะไม่มีที่ผิดพลาด คุณภาพของการแปลและความสามารถของผู้แปลเราก็ไม่รู้ ในขณะที่ การใช้คลังข้อมูลเทียบภาษา ทำให้เราเห็นการใช้ภาษาของผู้เชี่ยวชาญในสาขานั้นๆ และเป็นภาษาที่เป็นธรรมชาติเนื่องจากเป็นภาษาต้นฉบับทั้งคู่

ประโยชน์ของการใช้คลังข้อมูลภาษาเพื่อช่วยในการแปลเห็นได้ชัดในงานของ Zanettin (1998) ซึ่งได้กล่าวสรุปถึงประโยชน์ของการใช้คลังข้อมูลเทียบภาษาเพื่อการแปล ใน 3 ลักษณะ ดังนี้

1. คลังข้อมูลเทียบภาษาช่วยให้ผู้แปลใช้ภาษาแปลได้ดีมากขึ้น ตัวอย่างเช่น ในประโยคภาษาอิตาเลียน "*salire il gradino piu alto del podio*" (to climb onto the highest step of the podium) ผู้แปลจะแปลอย่างไร จะแปลแบบตรงตัว (literally) ได้ไหม เมื่อผู้แปลค้นคลังข้อมูลภาษาอิตาเลียนเรื่องกีฬาโอลิมปิค จะพบว่าสำนวน "*il gradino piu' alto del podio*" ใช้ในความหมายว่าชนะเลิศได้เหรียญทอง และเมื่อลองค้นคำว่า *podium* ในคลังข้อมูลภาษาอังกฤษที่เก็บจากหนังสือพิมพ์ The Independent และ The daily Telegraph ในส่วนที่เป็นเรื่องเดียวกัน พบว่ามีการใช้คำว่า *podium* ทำให้ยืนยันได้ว่าคำว่า *podium* สามารถใช้ในความหมายเดียวกับคำว่า *podio* ในภาษาอิตาเลียน แต่ในคลังข้อมูลนี้ไม่พบว่า *podium* เกิดร่วมกับ *the highest step* ในความหมายของการได้เหรียญทอง ดังนั้น หลักฐานจากคลังข้อมูลภาษาช่วยทำให้เห็นว่าการแปลแบบตรงตัวนั้นใช้ไม่ได้ เพราะจะทำให้คำแปลภาษาอังกฤษที่ได้ฟังดูแปลก

เนื่องจากไม่มีใครใช้กัน นอกจากนี้ คลังข้อมูลภาษาจะช่วยเห็นรูปแบบของคำที่เกิดร่วมกัน คือ *stand on the podium* ซึ่งพบ 6 ครั้งจากจำนวน 24 บรรทัดที่พบจากคลังข้อมูล ดังนั้น ผู้แปลอาจเลือกใช้สำนวนนี้เป็นคำแปลที่ต้องการ หรืออาจแปลง่ายๆว่า *won the gold medal* เลยก็ได้

2. คลังข้อมูลภาษาช่วยในการเรียนรู้เรื่องศัพท์ (terminology) และเนื้อหา (content) มากขึ้น คลังข้อมูลภาษาที่รวบรวมจากเอกสารเฉพาะด้านเป็นแหล่งข้อมูลที่สำคัญทางด้านศัพท์เฉพาะทางและเนื้อหา สามารถใช้คลังข้อมูลเทียบภาษาในชั้นเรียนเพื่อช่วยสนับสนุนสมมติฐานการแปล และหาคำแปลที่เหมาะสมในเอกสารเฉพาะด้าน Gavioli (1996) ได้ทดลองให้นักเรียนใช้คลังข้อมูลเทียบภาษาในการแปลเอกสารด้านการแพทย์ เพื่อหาความรู้ที่จำเป็นต่อการเข้าใจเอกสารต้นฉบับและทำการแปล ซึ่งเอกสารการแพทย์จะมีศัพท์เฉพาะทางจำนวนมากและมักมีที่มาจากภาษากรีกหรือลาติน ทำให้มีรูปคำที่คล้ายๆกันระหว่างภาษาอังกฤษและภาษาอิตาเลียน การใช้คลังข้อมูลเทียบภาษาช่วยให้ค้นหาคำในภาษาปลายทางได้โดยง่าย และทำให้เห็นคำที่มักปรากฏร่วมกันด้วย ซึ่งสามารถนำไปเทียบกับคำที่เกิดร่วมกันที่พบในภาษาต้นฉบับได้ เช่น *biopsia epatica* (ใน Zanettin 1998: 6) ดูไม่น่าจะเป็นปัญหาสำหรับการแปล เพราะมีคำในภาษาอังกฤษที่ตรงตัว คือ *biopsy* และ *hepatic* และทั้งสองคำนี้ก็เป็นศัพท์ทางการแพทย์ทั้งคู่ จึงน่าจะแปล *biopsia epatica* ว่า *hepatic biopsy* ได้ แต่ว่าเมื่อค้นคลังข้อมูลส่วนภาษาอังกฤษกลับไม่พบการใช้ *hepatic biopsy* แต่พบคำว่า *liver biopsy* ถึง 39 ครั้ง จึงน่าจะแปลว่า *liver biopsy* มากกว่า และเมื่อเปรียบเทียบระหว่าง *liver* (E) กับ *fegato* (I) และระหว่าง *hepatic* (E) กับ *epatico* (I) พบว่าคำคุณศัพท์ *hepatic* ในภาษาอังกฤษจะเกิดร่วมกับคำทั่วไป เช่น *disease lesion* หรือ *failure* ไม่เกิด

กับ *liver* แต่ในภาษาอิตาเลียน *epatico* มักจะเกิดร่วมกับ *fegato* *hepatic* จึงไม่น่าเป็นคำแปลของ *epatico* ในที่นี้

นอกจากนี้ Gavioli and Zanettin (1997) ได้รายงานการใช้ประโยชน์ของคลังข้อมูลภาษาในการแปล ตัวอย่างหนึ่งที่เขายกมาคือการแปลประโยคภาษาอิตาเลียน *In questo lavoro sono state esaminate 183 biopsie epatiche di etilisti con o senza marcatori HBV* ซึ่งในตอนแรกผู้แปลแปลว่า *In this paper 183 hepatic biopsies of alcoholics with or without HBV markers were examined* แต่เมื่อผู้แปลลองค้นหาโดยใช้คำค้นว่า biops*[1] พบ *liver biopsy/ies* 47 ครั้ง แต่ไม่พบ *hepatic biopsy* เลยแม้ว่าจะพบ *hepatic* ซึ่งเกิดกับคำอื่นๆ จึงเลือกแปลว่า *liver biopsies* แทน *hepatic biopsies* และเมื่อลองค้นคำว่า marker* พบว่ามีการใช้ *Patients without hepatitis C markers* 1 ครั้ง ซึ่งอาจทำให้คิดว่าที่ได้แปลไว้นั้นน่าจะใช้ได้ แต่เมื่อดูในข้อมูลที่ได้มา กลับพบ *positive/negative for HBV/HCV markers* เกิดบ่อยมากกว่า เมื่อค้นหา marker* HBV HCV ร่วมกับ positive หรือ negative ก็พบว่ามีการใช้ *negative for HBV/HCV markers* 5 ครั้ง และพบ *positive/negative for HCV/ HCV antibodies / HCV tests* 15 ครั้ง ผู้แปลจึงแก้ไขคำแปลใหม่เป็น *In this paper 183 liver biopsies of alcoholics who were positive or negative for HBV markers were examined*

ซึ่งเมื่อผู้แปลตรวจสอบในคลังข้อมูลภาษาต่อไป เพื่อตรวจสอบว่าการขึ้นต้นบทคัดย่อด้วย *in this paper* นั้นเหมาะสมไหม ซึ่งผลจากการค้น

---

[1] สัญลักษณ์ * ที่ใช้หมายถึงแทนตัวอักษรอะไรก็ได้ ในที่นี้จึงหมายถึงให้ค้นคำที่ขึ้นต้นด้วย biops

คลังข้อมูลที่เก็บบทคัดย่อพบว่าไม่มีการเขียนบทคัดย่อโดยขึ้นต้นด้วยคำว่า *in this paper* ในคลังข้อมูลเลย เมื่อตรวจดูการขึ้นต้นข้อความของบทคัดย่อ ก็พบการขึ้นต้นในสองลักษณะ แบบแรกเป็นการใช้ประโยค infinitive บอกจุดมุ่งหมาย แบบสองเป็นการใช้คำสรรพนาม *we* และตามด้วยคำกริยา เช่น *we study* ผู้แปลก็เลยเปลี่ยนคำแปลเป็น *We examined 183 liver biopsies of alcoholics who were positive or negative for HBV markers* ซึ่งเป็นคำแปลที่เหมาะสมมากกว่า

3. คลังข้อมูลเทียบภาษาสามารถใช้เพื่อศึกษาลักษณะของตัวบทประเภท (genre) หรือหัวข้อ (topic) ที่สนใจ ทำให้มีความรอบรู้ในเรื่องต่างๆมากขึ้น ซึ่งสามารถทำเป็นกิจกรรมก่อนหรือหลังการแปลก็ได้ คลังข้อมูลเทียบภาษาเป็นทรัพยากรที่ดีสำหรับใช้สำรวจภาษา ผู้เรียนสามารถใช้เพื่อทำความเข้าใจกับคำหรือวลีต่างๆ หรืออาจให้ผู้เรียนเปรียบเทียบศึกษาความคล้ายคลึงกันระหว่างภาษา ดูคำหรือวลีที่คล้ายคลึงกัน เช่น ดูการใช้ชื่อเฉพาะ (proper name) ในตัวอย่างการค้นชื่อว่า *Mitterrand* ในคลังข้อมูลข่าวต่างประเทศที่เกี่ยวข้องกับประเทศฝรั่งเศส ทั้งในภาษาอังกฤษและภาษาอิตาเลียน เมื่อแจกแจงรายการคำทั้งหมดที่ปรากฏร่วมกับ *Mitterrand* จะพบว่ามีรูปแบบการใช้ที่แตกต่างกันระหว่างสองภาษานี้ ในภาษาอิตาเลียน มีแนวโน้มที่จะเรียกชื่อเต็มหรือนามสกุล ได้เลย เช่น *Francois Mitterrand, Mitterrand* แต่ในภาษาอังกฤษจะใช้ *Presdient Mitterrand. President Francois Mitterrand, Mr Mitterrand* ซึ่งน่าประหลาดใจกับผลที่พบนี้ เพราะภาษาอิตาเลียนเป็นที่ทราบกันว่านิยมใช้คำนำหน้าชื่อ (title) มาก เช่น ใช้ *dottore* สำหรับคนที่จบปริญญา *professore* สำหรับอาจารย์มัธยม

จากนั้น อาจศึกษาต่อโดยดูบริบทต่างๆของคำว่า *President* และ

French ในภาษาอังกฤษ และ *presidente* และ *francese* ในภาษาอิตาเลียน ได้พบว่าในภาษาอังกฤษไม่ค่อยมีการใช้คำคุณศัพท์ *French* มากเท่าใด และมีแนวโน้มที่จะเลี่ยงการใช้ *the French President Francois Mitterrand* หรือ *the French President Mitterrand* แต่ในภาษาอิตาเลียนจะพบเป็นปกติ ขั้นต่อไปอาจสำรวจตามเนื้อหาทั้งหมด ซึ่งทำให้เห็นว่า ในภาษาอังกฤษใช้ *Mitterrand* เฉยๆ เฉพาะในหัวข้อข่าว (headline) ในเนื้อข่าวจะใช้ *president Mitterrand* และต่อๆ มาจะใช้ *Mr Mitterrand* หรือ *Mr President* ส่วนในภาษาอิตาเลียนจะเริ่มต้นใช้ *Francois Mitterrand* หรือ *il Presidente Mitterrand* และต่อมาจึงอ้างถึงโดยใช้ *il presidente / il presidente francese* หรือ *Mitterrand*

นอกจากนี้ อาจศึกษาต่อโดยดูความต่างของการใช้คำที่บอก direct speech เพราะสังเกตุเห็นว่าส่วนมากในข้อมูล *Mitterrand* จะปรากฏในลักษณะของการกล่าวหรือพูดถึงอะไรบางอย่าง เมื่อลองค้นหาคำสรรพนาม เช่น Mr, Mrs, Ms, และ Pr* ในบริบทที่มีเครื่องหมาย quotation จะทำให้เห็นความแตกต่างของการใช้ direct speech ในภาษาทั้งสอง ในภาษาอังกฤษ พบคำที่ใช้มากตามลำดับคือ *say, add, admit, annouce, ask, insist, declare, complain* ส่วนในภาษาอิตาเลียนพบคำว่า *dire, dichiarare, aggiungere, affermare, chiamare, concludere, insistere, ribadire* ซึ่งจะเห็นว่าในขณะที่ภาษาอังกฤษใช้คำ *say* กับ *add* มากกว่า 60% แต่คำที่เทียบเท่ากันในภาษาอิตาเลียน คือ *dire* กับ *aggiungere* กลับใช้ไม่ถึง 50% ซึ่งทำให้เห็นว่าคำที่เทียบเท่ากันระหว่างภาษานั้น อาจมีการใช้ที่แตกต่างกันได้ จึงสนใจศึกษาต่อโดยดูคำกริยาที่ใช้ในภาษาอังกฤษและอิตาเลียนที่เป็นคำที่คู่เสมือนหรือเป็นคำร่วมเชื้อสาย (cognate) เช่น *affermare* (I) ในพจนานุกรมสองภาษาจะให้คำแปลเป็นภาษาอังกฤษว่า *affirm, assert, state* ซึ่งเมื่อค้นในคลังข้อมูลเทียบภาษาจะ

พบ *affermare* อยู่จำนวนหนึ่ง แต่จะไม่พบคำว่า *affirm, assert, state* เลย จึงสรุปได้ว่า เราอาจใช้ *affermare* แปลคำว่า *say* และ *declare* และถ้าจะแปล *affermare* จากภาษาอิตาเลียนเป็นภาษาอังกฤษก็ไม่น่าจะเลือกแปลเป็นคำ *affirm* หรือ *assert*

ตัวอย่างที่ชัดเจนอีกตัวอย่างหนึ่ง คือคำว่า *prices* กับ *prezzi* ซึ่งดูไม่น่ามีปัญหาในการใช้ เพราะดูเป็นคำแปลที่เทียบเท่ากัน (translation equivalent) แต่เมื่อค้นดูจากคลังข้อมูลเทียบภาษาจะพบความแตกต่างกัน เช่น *prezzi* มักนำหน้าด้วยคำชี้เฉพาะ (definite article) *i* แต่ *prices* ไม่มีคำชี้เฉพาะนำหน้า หรือเมื่อพิจารณาดูคำปรากฏร่วมของคำทั้งสองคำนี้ก็จะพบว่ามีความแตกต่างกัน สองคำนี้จึงมีการใช้ที่แตกต่างกัน

นอกจากประโยชน์ของการใช้คลังข้อมูลภาษาเพื่อเป็นทรัพยากรการแปลตามที่ Zanettin (1998) กล่าวสรุปไว้ การทดลองที่จัดขึ้นในมหาวิทยาลัยต่างๆ จำนวนหนึ่งยังสามารถยืนยันถึงประโยชน์ของการใช้คลังข้อมูลภาษาเป็นทรัพยากร ดังตัวอย่างต่อไปนี้

8.2.1 การทดลองที่มหาวิทยาลัย Dublin City

Bowker (1998) รายงานการทดลองที่ทำขึ้นที่ School of Applied Language and Intercultural Studies มหาวิทยาลัย Dublin City เพื่อศึกษาประโยชน์ของการใช้คลังข้อมูลภาษาเฉพาะทาง (specialized monolingual corpora) โดยมีสมมติฐานว่าคลังข้อมูลเฉพาะทางของภาษาต้นฉบับ (ที่เป็นภาษา L1) เป็นทรัพยากรที่มีประโยชน์ต่อการแปลจาก L2 เป็น L1 กล่าวคือ คลังข้อมูลภาษาเดียวมีประโยชน์ต่อการเสริมสร้างความเข้าใจในสาขามากกว่าทรัพยากรแบบดั้งเดิม เช่น วารสาร พจนานุกรมภาษาเดียว อย่างน้อยในสองลักษณะ

1. ช่วยให้เข้าใจในเนื้อหาสาขานั้นได้เร็วกว่าและถูกต้องมากกว่า ทำให้การแปลผิดเนื่องจากความเข้าใจผิดพลาดน้อยกว่า ซึ่งการรวบรวมเอกสารในสาขานั้น ผู้แปลอาจต้องไปรวบรวมจากหลายที่ จากห้องสมุดต่างๆ เป็นการเสียเวลาและยังได้เอกสารที่ต้องการมาเพียงจำนวนหนึ่งเท่านั้น การค้นหาคำอธิบายในเรื่องที่ต้องการก็ทำได้ยาก ต้องอาศัยดัชนีที่ทำไว้หรืออ่านจากสารบัญ อ่านจากเรื่องจริงๆ ในขณะที่การใช้คลังข้อมูลภาษานั้นสะดวกรวดเร็วกว่า ผู้ใช้สามารถเข้าถึงข้อมูลที่ต้องการโดยการใช้โปรแกรมค้นหาคำและบริบทที่ต้องการได้ และยังเก็บไว้ใช้ได้กับการแปลครั้งต่อๆไปด้วย

2. ช่วยให้เลือกใช้การแปลได้อย่างถูกต้อง ทำให้เลือกใช้ศัพท์เฉพาะทางได้อย่างถูกต้อง และรู้ว่าคำใดควรใช้ร่วมกับคำใด มีผู้เปรียบว่าการศึกษาคำโดยแยกออกจากบริบทก็เหมือนการศึกษาสัตว์ที่ถูกสตาฟไว้ การจะศึกษาสัตว์นั้นจำเป็นต้องศึกษาการใช้ชีวิตความเป็นอยู่ตามธรรมชาติ การศึกษาคำก็ต้องพิจารณาการใช้คำตามธรรมชาติที่เกิดขึ้น (Roumen and van der Ster 1993: 215-216) แต่การดูรูปแบบ (pattern) ของการใช้ด้วยตาเปล่านั้นยากเพราะคำที่ต้องการศึกษาอาจปรากฏอยู่ในที่ต่าง ๆ ห่างกันหลายสิบหลายร้อยหน้า แต่ถ้าใช้โปรแกรมคอนคอร์แด๊นซ์ โปรแกรมจะช่วยดึงรูปแบบการใช้ออกมาให้เราเห็นติดๆกันได้ ทำให้เห็นภาพที่ชัดเจนกว่า นอกจากนี้คอมพิวเตอร์ยังบอกถึงค่าสถิติต่างๆของการใช้คำและคำที่ปรากฏร่วมกัน

การทดลองนี้ใช้นักศึกษาสาขาศาสตร์การแปล 14 คนเป็นกลุ่มตัวอย่าง แต่เนื่องจากต้องการทดสอบวัดประสิทธิภาพของการใช้ประโยชน์คลังข้อมูลภาษาเพื่อการแปล จึงไม่สามารถใช้นักศึกษาที่เป็นนักแปลมือใหม่ จึงต้องเลือกนักศึกษาที่มีชั่วโมงฝึกหัดการแปลมาแล้วพอสมควร และต้องเป็นผู้พูดภาษาอังกฤษเพราะคลังข้อมูลที่ให้ใช้เป็นคลังข้อมูลภาษาอังกฤษ และเนื่องจาก

การทดลองไม่เกี่ยวข้องกับการปรับปรุงประสิทธิภาพของภาษาต่างประเทศ นักศึกษาจึงต้องเป็นผู้ที่มีความรู้ภาษาต่างประเทศดีพอที่จะอ่านต้นฉบับได้ดี แม้ว่านักศึกษาจะไม่ใช่ผู้เชี่ยวชาญในสาขาที่แปล แต่ก็ควรจะมีทัศนคติที่ดีต่อการแปลในสาขานั้น และเนื่องจากเป็นการทดลองที่ต้องใช้โปรแกรมคอนคอร์แด๊นซ์กับคลังข้อมูลภาษา นักศึกษาจึงต้องมีความคุ้นเคยกับการใช้คอมพิวเตอร์พอสมควร

จากกฎเกณฑ์เหล่านี้ จึงได้เลือกนักศึกษาที่ศึกษา Applied Computational Linguistics ที่ DCU จำนวน 14 คน เพราะนักศึกษาเหล่านี้เป็นผู้พูดภาษาอังกฤษ อยู่ปีสุดท้ายในในระดับปริญญาตรี และได้เรียนวิชาการแปลมาแล้ว อย่างน้อย 3 หน่วยกิต เรียนฝรั่งเศสมาอย่างน้อย 6 หน่วยกิตและเคยไปอยู่ในฝรั่งเศสหรือเบลเยี่ยมมาหนึ่งปี ทุกคนคุ้นเคยกับคอมพิวเตอร์แต่ไม่มีผู้ใดเชี่ยวชาญเรื่อง optical scanning technology ทุกคนเคยเรียนวิชาภาษาศาสตร์คลังข้อมูล เคยใช้คลังข้อมูลภาษาและโปรแกรมคอนคอร์แด๊นซ์แล้ว

ตัวบทที่ให้นักศึกษาแปลมาจากเอกสารที่เป็นเรื่องเฉพาะสาขาแต่ก็ไม่ใช่เรื่องทางเทคนิคมากๆ เป็นเรื่องที่ผู้วิจัยเองก็มีความรู้พอที่จะประเมินผลการแปลได้ และเป็นตัวบทที่ไม่ยาวมากเพราะผู้เข้าทดสอบเป็นอาสาสมัคร ตัวบทที่ใช้จึงตั้งไว้ว่ายาวไม่เกิน 300 คำ จึงได้เลือกตัวบท 2 ชิ้นจากวารสารวิทยาศาสตร์ทั่วไปที่เป็นเรื่องของ optical scanner ในภาษาฝรั่งเศส

ทรัพยากรที่กำหนดให้นักศึกษาใช้ได้ มีดังนี้

1. ทรัพยากรแบบดั้งเดิมประกอบด้วย

- พจนานุกรมต่างๆ ได้แก่ Oxford English Dictionary และ พจนานุกรมทางคอมพิวเตอร์ 5 เล่ม

- แหล่งข้อมูลอื่นที่ไม่ใช่พจนานุกรม ได้แก่ สารานุกรม (encyclopaedia) ด้านอิเล็กทรอนิกส์ หนังสือคู่มือการใช้ของเครื่องสแกนเนอร์ บทความเกี่ยวกับเครื่องสแกนเนอร์และโปรแกรมรู้จำตัวอักษร (OCR) ในวารสารทั่วไปทางคอมพิวเตอร์ เอกสาร (monograph) เรื่อง desktop publishing

2. คลังข้อมูลภาษาประกอบด้วย คลังข้อมูลภาษาที่รวบรวมตัวบทจากในซีดีรอมชื่อ Computer Select แม้คลังข้อมูลที่สร้างนี้จะไม่ใช่คลังข้อมูลที่มีข้อมูลครบถ้วนทุกด้านแต่ก็มีประโยชน์ให้ข้อมูลที่น่าสนใจได้ เนื่องจากผู้วิจัยเคยทำศัพท์เฉพาะทางเกี่ยวกับสแกนเนอร์มาก่อน จึงใช้รายการศัพท์ที่มีเป็นคำค้นในการค้นหาบทความในซีดีรอมนั้น โดยรวบรวมบทความจากซีดีรอมครอบคลุมช่วงพฤษภาคม 1989 ถึงกุมภาพันธ 1995 คลังข้อมูลที่สร้างขึ้นนี้มีขนาด 10 ล้านคำ ในส่วนของโปรแกรมคอนคอร์แด๊นซ์ที่ใช้คือโปรแกรม Word Smith

วิธีการทดลองที่ทำได้แบ่งข้อมูลเป็นสองส่วน ให้นักศึกษาแปลโดยใช้ทรัพยากรทั้งสองแบบ แบบแรกประกอบด้วยพจนานุกรมสองภาษา (bilingual dictionary) พจนานุกรมเฉพาะทางภาษาเดียว (monolingual specialized lexicographic) และทรัพยากรอื่นๆที่ไม่ใช่พจนานุกรม (non-lexicographic resource) แบบที่สองประกอบด้วยพจนานุกรมสองภาษา และคลังข้อมูลเฉพาะทาง (specialized monolingual corpus) และโปรแกรม Word Smith และเพื่อมิให้มีข้อสงสัยเรื่องของความต่างของเอกสารและความต่างของนักเรียน เลยแบ่งนักศึกษาเป็นสองกลุ่ม กลุ่มแรกแปลตัวบทที่หนึ่งโดยใช้ทรัพยากรชุดแรกก่อน กลุ่มหลังแปลตัวบทที่สองโดยใช้ทรัพยากรชุดแรกก่อน โดยกำหนดเวลาให้ 2 ชั่วโมงสำหรับแปลตัวบททั้งสองชิ้น และให้นักศึกษาเขียนวิจารณ์การใช้ทรัพยากรทั้งสองแบบ

ผลการวิเคราะห์ทำในรูปของการนับจำนวนข้อผิดพลาดทั้งที่เป็นข้อผิดพลาดด้านความเข้าใจเนื้อหา และข้อผิดพลาดด้านการเลือกใช้คำแปลผิด (รวมถึงการเลือกใช้คำผิด ใช้วลีที่ไม่ใช่สำนวนปกติ การเขียนผิดไวยากรณ์ และใช้ภาษาไม่เหมาะสมกับทำเนียบภาษา) ผลการวิเคราะห์สรุปประโยชน์ของการใช้คลังข้อมูลภาษาได้เป็นสามแนวทาง

1. ช่วยทำให้เข้าใจเนื้อหาได้ดีขึ้น

เมื่อเปรียบเทียบ การแปลแบบใช้ทรัพยากรแบบดั้งเดิมกับการแปลแบบใช้คลังข้อมูลภาษาเฉพาะทางที่เป็นภาษาแม่ (specialized monolingual native-language corpus) เช่น ในตัวบทที่หนึ่ง นักเรียนส่วนมากมีปัญหาไม่เข้าใจความหมายของประโยคแรก (ในส่วนที่ขีดเส้นใต้) *quelle que soit leur sensibilité aux nuances, leur rapidité, leur précision,...* (Bowker 1998: 10) ซึ่งควรแปลได้ข้อความ *Regardless of such characteristics as colour-recognition capability, speed, precision,...* แต่พจนานุกรมเฉพาะทางไม่ได้พูดเรื่อง color-recognition cabability ไว้ ในหนังสือคู่มือการใช้สแกนเนอร์บอกให้ไปดูที่คำว่า *black-and-white vs. color scanners* ส่วน desktop publishing monograph มีกล่าวถึงความแตกต่างระหว่าง non-greyscale, grayscale, และ color scanner อยู่ครึ่งหน้า แต่ไม่ได้ทำดัชนีไว้ท้ายเล่มให้ค้นจึงค้นพบได้ยาก สำหรับบทความในวารสารมีการกล่าวถึงเรื่องนี้ไว้อย่างละเอียดแต่ไปอยู่ที่ตอนท้ายสุดของบทความและไม่ได้แยกให้เห็นต่างจากส่วนอื่นๆ ชัดเจนจึงทำให้หาพบได้ยากเช่นกัน ดังนั้น จึงไม่น่าประหลาดใจที่ไม่มีนักศึกษาคนใดที่ใช้ทรัพยากรแบบดั้งเดิมสามารถแปลประโยคนี้ได้ถูกต้อง ส่วนพวกที่ใช้คลังข้อมูลภาษา สามารถเข้าถึงตัวเอกสารที่เกี่ยวข้องได้โดยตรงมากกว่า จึงสามารถอ่านและทำความเข้าใจเนื้อหาได้ดีกว่า แม้ว่านักศึกษาอาจไม่เข้าใจ

ความหมายคำว่า *nuances* แต่จากการค้นหาคำว่า *sensitiv** จะพบว่าเกิดร่วมกับคำต่างๆ เช่น เกิดร่วมกับ *colour* 5 ครั้ง เกิดร่วมกับ *greyscale* 4 ครั้ง เกิดร่วมกับ *shade** 122 ครั้ง เกิดร่วมกับ *shading* 8 ครั้ง นักศึกษาบางคนจึงเดาว่าน่าจะแปลคำว่า *nuances* ว่า *shade* และมีหนึ่งคนที่เลือกใช้คำว่า colour แม้ว่าจะไม่มีใครแปลได้ถูกต้อง แต่จากคำแปลที่ได้จะเห็นว่าในกลุ่มนักศึกษาที่ใช้คลังข้อมูลภาษานั้น 4 คนจากทั้งหมด 7 คน มีความเข้าใจในเรื่องนี้มากกว่า

2. ช่วยให้เลือกใช้คำแปลได้ถูกต้องเหมาะสม

ในตัวบทหมายเลขหนึ่ง คำว่า *vitre* สามารถแปลว่า *glass platen* หรือ *scan bed* ก็ได้ แต่ในพจนานุกรมเฉพาะทางไม่มีคำว่า *glass platen* และคำแปลที่ได้จากพจนานุกรมสองภาษามีแต่คำว่า *glass*, *pane of glass*, *window* คำว่า *glass pattern* นี้มีปรากฏอยู่ในหนังสือคู่มือการใช้ แต่ไม่มีในดัชนีท้ายเล่มให้ค้น นักศึกษาจึงหาพบได้ยาก ส่วนใน monograph ของ desktop publishing ไม่มีคำนี้อยู่ ดังนั้น นักศึกษาที่ใช้ทรัพยากรแบบดั้งเดิม ไม่มีใครเลือกคำแปลได้ถูกต้อง ในขณะที่นักศึกษาที่ใช้คลังข้อมูลเทียบบทแปลได้ถูก 3 คน เพราะสามารถพบคำว่า *glass pattern* ซึ่งมีความถี่สูงในลำดับต้นๆของคำที่เกิดร่วมกับ *glass*

อีกตัวอย่างหนึ่งคือคำว่า *scanner a plat* ซึ่งต้องแปลว่า *flatbed scanner* แต่นักศึกษาที่ใช้ทรัพยากรแบบดั้งเดิมแปลว่า *flat scanner* แม้ว่าคำว่า *flatbed scanner* จะมีปรากฏใช้อยู่ใน monograph ของ desktop publishing หลายครั้ง ส่วนนักศึกษาที่ใช้คลังข้อมูลเทียบบทแปลได้ถูกต้องทุกคน บางคนให้ข้อมูลเพิ่มเติมด้วยว่าพบทั้งคำว่า *flatbed* กับ *flat-bed* แต่คำแรกเกิดมากกว่าคือปรากฏ 1508 ครั้ง ส่วนคำหลัง ปรากฏ 92 ครั้ง จึงเลือกใช้คำแรกเป็นคำแปล

3. ช่วยให้เลือกใช้สำนวนภาษาที่เหมาะสม (improve idiomatic construction)

ในตัวอย่าง *photodiodes sensibles a la lumiere* พจนานุกรมเฉพาะทางมีคำว่า *photodiode* ใน monograph ของ desktop publishing แม้จะมีการใช้คำว่า *light-sensitive elements* แต่ไม่ปรากฏในดัชนีท้ายเล่มสำหรับค้นจึงหาได้ยาก ส่วนในบทความวารสารไม่มีกล่าวถึง *photodiode* เลย ดังนั้น ในกลุ่มนักศึกษาที่ใช้ทรัพยากรแบบดั้งเดิมนี้เมื่อเห็นว่าในพจนานุกรมมีคำว่า *photodiode* จึงเลือกแปลโดยใช้คำนี้และคงโครงสร้างแบบเดิมเป็น *photodiodes sensitive to the light* แม้ว่าโครงสร้างนี้จะถูกไวยากรณ์แต่ไม่ใช่สำนวนที่ใช้กันในสาขา ส่วนในกลุ่มที่ใช้คลังข้อมูลภาษาเนื่องจากเห็นจากคลังข้อมูลว่าไม่มีการแปลแบบนี้ นักศึกษาทุกคนจึงเลือกแปลว่า *light-sensitive photodiodes* หรือไม่ก็ *photosensitive diodes* ตามที่พบในคลังข้อมูลทั้งสองแบบ

อีกตัวอย่างหนึ่ง คือ ***la tête de numéerisation du scanner*** ซึ่งควรจะแปลว่า *scan head* แต่อาจแปลเป็น *scanning head* หรือ *scanner head* ก็ได้ แต่ในกลุ่มที่ใช้ทรัพยากรแบบดั้งเดิม พบว่านักศึกษา 3 คนเลือกใช้โครงสร้างตามแบบต้นฉบับเป็น *head of the scanner* แต่ในกลุ่มที่ใช้คลังข้อมูลภาษาไม่มีผู้ใดแปลโดยใช้โครงสร้างแบบนี้เลย

ผลจากการทดลองครั้งนี้ สามารถสรุปได้ว่า ผลงานแปลของนักศึกษาที่ใช้คลังข้อมูลภาษา จะมีข้อผิดพลาดน้อยกว่าของคนที่ใช้ทรัพยากรแบบดั้งเดิม เมื่อพิจารณาดูในรายละเอียด ส่วนของข้อผิดพลาดจากความเข้าใจ ส่วนของการเลือกใช้ศัพท์ และส่วนของการใช้โครงสร้างที่ไม่ใช่สำนวนภาษาที่เหมาะสม (non-idiom construction) จะพบว่าการใช้คลังข้อมูลภาษาช่วยเสริมให้ผู้ใช้แปลงานได้ดีขึ้น ส่วนข้อผิดพลาดทางไวยากรณ์ (grammatical error) เป็นปัญหาจากความสับเพร่าของผู้แปลเอง ในส่วนของการใช้ภาษาผิดทำเนียบภาษา (incorrect register) ก็ดูเป็นปัญหาส่วนบุคคลมากกว่าที่จะเกี่ยวกับการใช้หรือ

การไม่ใช้คลังข้อมูลภาษา

สำหรับข้อคิดเห็นจากนักศึกษาผู้เข้าร่วมในการทดลอง นักศึกษาที่ร่วมในการทดลองมีความรู้สึกว่าทรัพยากรแบบดั้งเดิมนั้นไม่เพียงพอต่อการใช้แปล ไม่สามารถหาข้อมูลที่ต้องการ เช่น ข้อมูลเกี่ยวกับสแกนเนอร์แบบต่างๆ แต่นักศึกษา 2 คนก็บอกว่าพอใจที่จะใช้ทรัพยากรแบบเดิมเพราะสะดวกและใช้ง่าย นักศึกษา 4 คนบอกว่าการใช้คลังข้อมูลภาษานั้นทำให้เสียเวลามาก แต่ก็บอกว่าเมื่อคุ้นเคยกับระบบแล้ว คงไม่เสียเวลามากนัก ในกลุ่มนักศึกษาที่บอกว่าคลังข้อมูลภาษามีประโยชน์ต่อการแปลนั้น ได้ชี้ให้เห็นประโยชน์ในประเด็นเหล่านี้

- คลังข้อมูลภาษาช่วยให้เข้าใจเนื้อหาสาขาได้ดีขึ้น ช่วยให้เข้าใจความหมายของคำได้ดีขึ้น
- คลังข้อมูลภาษาช่วยให้เลือกใช้คำแปลได้เหมาะสมมากขึ้น
- ความถี่ที่เห็นจากคลังข้อมูลภาษา ช่วยให้ตัดสินใจได้ว่าในคำแปลหลายๆ คำที่เป็นไปได้ ควรเลือกคำที่มีความถี่มากที่สุดก่อน
- คลังข้อมูลภาษาช่วยยืนยันว่าคำแปลที่ต้องการใช้นั้นมีใช้กันอยู่จริงๆ

โดยสรุป การทดลองครั้งนี้ได้สนับสนุนสมมติฐานที่ว่าคลังข้อมูลภาษาเฉพาะทางที่เป็นภาษาแม่นั้นเป็นทรัพยากรที่มีประโยชน์ต่อนักแปลแม้ว่าจะเป็นการแปลเป็นภาษาแม่ของผู้แปลเองก็ตาม

8.2.2 การทดลองที่มหาวิทยาลัย Innsbruck

Friedbichler (1997) ทำการทดลองเพื่อศึกษาว่า พจนานุกรม คลังข้อมูลภาษา และอินเทอร์เน็ต (เว็บ) จะเป็นประโยชน์ต่อการแปลอย่างไร โดยมีสมมติฐานว่าคลังข้อมูลภาษาเป็นทรัพยากรที่มีประโยชน์มากเป็นตัวเชื่อมต่อระหว่างพจนานุกรมและอินเทอร์เน็ตได้ดี การทดลองนี้ใช้กลุ่มตัวอย่าง

คือนักศึกษา 24 คนที่ลงเรียนในรายวิชา medical translation วิธีทดลองคือให้นักศึกษาสามารถใช้พจนานุกรมทางแพทย์ในรูปอิเล็กทรอนิกส์ คลังข้อมูลเทียบภาษาทางด้าน neurological โปรแกรมคอนคอร์แด๊นซ์ และอินเทอร์เน็ตเป็นทรัพยากรเพื่อช่วยในการแปล จากนั้นให้นักเรียนประเมินผลการใช้งานทรัพยากรต่างๆที่ได้ใช้

ผลที่สรุปได้จากการทดลองครั้งนี้ พบว่าคลังข้อมูลภาษาเป็นทรัพยากรที่นักศึกาาประทับใจมากที่สุดและมีประโยชน์มาก ส่วนอินเทอร์เน็ตเป็นประโยชน์สำหรับการค้นหาข้อมูลบางอย่างที่เฉพาะเจาะจง แต่ก็เป็นการใช้ที่ค่อนข้างเปลืองเวลามาก ส่วนพจนานุกรมทางแพทย์ที่อยู่ในรูปของอิเล็กทรอนิกส์นั้นก็มีประโยชน์มาก โดยเฉพาะอย่างยิ่งการที่สามารถค้นหาคำต่างๆได้ สำหรับพจนานุกรมทั่วๆไป ข้อดีคือเป็นที่คุ้นเคยต่อการใช้

ในกลุ่มนักศึกษาที่คุ้นกับการใช้คอมพิวเตอร์ จะเห็นประโยชน์จากการใช้คลังข้อมูลภาษาชัดเจนมาก แต่ในกลุ่มนักศึกษาที่ไม่คุ้นกับการใช้คอมพิวเตอร์ การใช้คอมพิวเตอร์กลับเป็นอุปสรรคเสียเอง แต่นักศึกษาก็รู้สึกประทับใจกับสิ่งที่เห็น และว่าการจะได้ประโยชน์จริง อาจต้องให้ผ่านช่วงการเรียนรู้ไปเสียก่อนจึงจะใช้งานคลังข้อมูลภาษาได้อย่างมีประสิทธิภาพ ส่วนการใช้อินเทอร์เน็ตหรือเว็บแทนคลังข้อมูลภาษานั้นเสียเวลามากกว่าเป็น 10 เท่า เพราะไม่สามารถจัดเรียงรายการที่ค้นได้เหมือนเวลาที่ใช้โปรแกรมคอนคอร์แด๊นซ์ ยังคงต้องการตัวช่วยค้น (search engine) ที่สามารถจัดเรียงข้อมูลได้แบบโปรแกรมคอนคอร์แด๊นซ์อยู่

นอกจากการทดลองที่กล่าวมาแล้ว ได้มีการทดลองในลักษณะคล้ายๆกันที่ School for Translators and Interpreters มหาวิทยาลัย Bologna (Gavioli and Zanettin, 1997) และที่ Department of Translation Studied

มหาวิทยาลัย Tempere (Varantola, 1997) การทดลองเหล่านี้ต่างก็สนับสนุนว่าคลังข้อมูลภาษาเป็นทรัพยากรที่เป็นประโยชน์ต่อการแปลอย่างยิ่ง แม้คลังข้อมูลภาษาที่ใช้จะไม่ใช่คลังข้อมูลเทียบบท แต่เป็นคลังข้อมูลเทียบภาษา หรือคลังข้อมูลภาษาเดียว ซึ่งอาจเป็นคลังข้อมูลเฉพาะทางที่จัดรวบรวมขึ้นเอง ก็เป็นประโยชน์อย่างมากต่อการทำความเข้าใจและการเลือกคำแปลที่เหมาะสม

บทที่ 5 คลังข้อมูลกับการทำพจนานุกรม

1. บทนำ

การทำพจนานุกรมจัดว่าเป็นงานหลักประเภทหนึ่งที่ได้ใช้ประโยชน์จากคลังข้อมูลภาษาอย่างมาก งานส่วนที่เกี่ยวข้องกับการทำพจนานุกรม มีตั้งแต่เรื่องการคัดเลือกข้อมูลภาษาที่ควรนำมาไว้ในคลังข้อมูลภาษา การสกัดข้อมูลที่เกี่ยวข้องเพื่อใช้ในการวิเคราะห์ศัพท์ ตลอดจนถึงการใช้เครื่องมือโปรแกรมต่างๆ เพื่อช่วยให้ผู้ทำพจนานุกรมทำงานกับคลังข้อมูลภาษาและใช้ประโยชน์จากคลังข้อมูลภาษาได้มากที่สุด

การทำพจนานุกรมแต่เดิมนั้น เช่น การทำ Oxford English Dictionary เป็นงานที่ต้องอาศัยข้อมูลตัวอย่างประโยคที่จดบันทึกไว้ในบัตรรายการคำจำนวนมากเพื่อทำความเข้าใจและเขียนนิยามความหมายของคำแต่ละคำ จนกระทั่งในช่วงทศวรรษที่ 1980 ที่มีโครงการ COBUILD เกิดขึ้นและได้ใช้ประโยชน์จากคลังข้อมูลภาษาขนาดใหญ่ที่เก็บในคอมพิวเตอร์ (Killgariff 2003) นับจากนั้นมาพจนานุกรมต่างๆ ก็อาศัยคลังข้อมูลภาษาขนาดใหญ่เป็นพื้นฐานในการทำงาน ตัวอย่างเช่น พจนานุกรมกลุ่ม Cobuild ได้อาศัย Bank of English ซึ่งเป็นคลังข้อมูลภาษาแบบสมดุลขนาด 650 ล้านคำ และอาศัย Collins Corpus ขนาด 2.5 พันล้านคำ พจนานุกรมกลุ่ม Longman อาศัยคลังข้อมูลภาษา Longman Corpus Network ที่มีขนาด 330 ล้านคำ พจนานุกรมของ Macmillion อาศัยคลังข้อมูล World English Corpus ขนาด 200 ล้านคำ และพจนานุกรมกลุ่ม Cambridge ที่ได้อาศัยคลังข้อมูล Cambridge International Corpus ขนาด 1

พันล้านคำ[1] พจนานุกรมที่อ้างอิงคลังข้อมูลภาษาเหล่านี้มักชูจุดเด่นของตนว่าครอบคลุมการใช้ภาษาที่แท้จริงและทันสมัย

2. พจนวิทยาและการทำพจนานุกรม

พจนวิทยา (lexicology) เป็นการศึกษาเรื่องของคำ ซึ่งคำถามพื้นฐานเลยก็เป็นเรื่องว่า คำคืออะไร ขอบเขตคำกำหนดอย่างไร เช่น *lunchtime, lunch-time, lunch time* ควรเป็นหนึ่งคำหรือสองคำ รวมไปถึงกรณีคำซึ่งมีรูปหรือเสียงเหมือนกันซึ่งก็จะพิจารณาว่าเป็นคนละคำถ้ามีความหมายต่างกัน แต่ก็มีกรณีที่คำมีความหมายใกล้กัน เช่น *part* 'portion' กับ *part* 'to separate' จะจัดเป็นคนละคำหรือไม่ ปัญหาเหล่านี้ยิ่งมีมากขึ้นในภาษาที่ไม่มีการเขียนแยกคำอย่างชัดเจนแบบภาษาอังกฤษ เช่น ภาษาไทย ภาษาจีน การศึกษาเรื่องคำยังครอบคลุมแง่มุมต่างๆ ได้แก่ ประวัติความเป็นมา (etymology) อรรถศาสตร์คำศัพท์ (lexical semantics) ประเภทของคำ การสร้างคำ เป็นต้น

ในแต่ละภาษาจะมีชุดของรายการคำ (vocabulary) หรือคลังคำ (lexicon) ที่ใช้ในภาษานั้น ซึ่งสามารถแยกเป็นสองกลุ่ม คือ คำเนื้อหา (content words) ซึ่งเป็นคำใช้แทนสิ่งต่างๆ ทั่วไป เช่น *ภูเขา ทะเล รัก เกลียด* และคำหน้าที่ (function words) ซึ่งเป็นกลุ่มคำปิดที่มีหน้าที่ทางไวยากรณ์มากกว่าแสดงเนื้อหาความหมาย เช่น *บน ใน ของ จะ* พจนวิทยาเป็นการศึกษาที่เน้นเรื่องของคำเนื้อหาเป็นหลัก ซึ่งวิธีที่ใช้อธิบายคำในภาษา คือ การอธิบายในรูปของพจนานุกรม (dictionary) และอรรถาภิธาน (thesaurus)

คำว่า dictionary เป็นคำที่มาจากภาษาลาติน dictionarius ที่ใช้ในปี ค.ศ. 1225 เป็นชื่อหนังสือรวบรวมศัพท์ในภาษาลาติน ส่วนคำภาษาอังกฤษ

---

[1] ข้อมูลจากอินเทอร์เน็ตเมื่อเดือนมิถุนายน พ.ศ.2553

dictionary พบหลักฐานใช้ครั้งแรกในปีค.ศ.1526 (Béjoint 1994:6) พจนานุกรมถูกมองว่าเป็นคู่ตรงข้ามกับหนังสือไวยากรณ์ โดยมองว่าพจนานุกรมจะให้ข้อมูลเกี่ยวกับคำ ในขณะที่หนังสือไวยากรณ์จะให้ข้อมูลเกี่ยวกับการประกอบคำเป็นหน่วยที่ใหญ่ขึ้น อย่างไรก็ตาม พจนานุกรมและไวยากรณ์ก็ไม่ใช่สิ่งที่แยกขาดจากกันได้ชัดเจน และพจนานุกรมปัจจุบันมีการให้ข้อมูลทางไวยากรณ์มากขึ้น โดยเฉพาะพจนานุกรมที่ทำขึ้นสำหรับผู้เรียนชาวต่างประเทศ เช่น พจนานุกรม Longman Dictionary of Contemporary English เป็นพจนานุกรมที่เรียกว่าพจนานุกรมผสมไวยากรณ์ (dictionary-cum-grammar)

อย่างไรก็ตาม สิ่งที่เรียกได้ว่าเป็นพจนานุกรมนั้นมีมาช้านานแล้ว ตั้งแต่ยุคที่ชาวบาบิโลนจดบันทึกคำภาษาสุมาเรียนในดินเหนียวเป็นหมวดหมู่คล้ายกับอรรถาภิธาน จึงถือได้ว่าเป็นพจนานุกรมสองภาษาฉบับแรกก็ได้ ในส่วนของอารยธรรมอื่นๆ เช่น จีน อียิปต์ กรีก โรมัน ต่างก็มีการรวบรวมความรู้เป็นพจนานุกรมภาษาของตนเองขึ้นมา ในยุคต่อมา การทำพจนานุกรมมักเป็นไปเพื่อประโยชน์ในการศึกษาศาสนา มีการรวบรวมพจนานุกรมภาษาลาติน ภาษาสันสกฤต ภาษาจีน (สำหรับศึกษาปรัชญาขงจื่อ) ภาษาอารบิค (สำหรับคัมภีร์โกราน) (Sterkenburg 2003:8-9)

พจนานุกรมสามารถแยกประเภทเป็นพจนานุกรมทั่วไปหรือพจนานุกรมเฉพาะทาง (general and specialized dictionaries) หากมองจากโครงสร้างใหญ่ (macro structure) พจนานุกรมทั่วไปจะครอบคลุมคำในคลังคำ ส่วนพจนานุกรมเฉพาะทางจะครอบคลุมคำส่วนหนึ่งอาจเป็นคำในแวดวงจำกัด เช่น เป็นเฉพาะคำแสลง เป็นต้น หากมองจากโครงสร้างย่อย (micro structure) พจนานุกรมแบบทั่วไปจะให้ข้อมูลคำครบถ้วนกว่า มีคำนิยาม คำอ่าน หมวดคำ ตัวอย่างการใช้ ประวัติความเป็นมา เป็นต้น แต่ถ้าให้ข้อมูลเฉพาะส่วนก็เป็น

พจนานุกรมเฉพาะ เช่น พจนานุกรมคำพ้อง พจนานุกรมนิรุกติศาสตร์ (dictionary of etymology)

พจนานุกรมอาจแยกประเภทตามภาษาเป็นพจนานุกรมภาษาเดียวหรือสองภาษา (monolingual and bilingual dictionaries) โดยพจนานุกรมสองภาษาจะให้ข้อมูลคำในภาษาที่สองที่มีความหมายเทียบเท่ากับคำในภาษาที่หนึ่ง หรืออาจแยกประเภทตามผู้ใช้เป้าหมายว่าเป็นพจนานุกรมสำหรับผู้เรียนชาวต่างประเทศหรือสำหรับเจ้าของภาษา (foreign learners' and native speakers' dictionaries) พจนานุกรมภาษาเดียวสำหรับผู้เรียนชาวต่างประเทศ เช่น Collin Cobuild English Language Dictionary จะเลือกบรรจุคำที่มักใช้บ่อยเพราะเป็นประโยชน์สำหรับผู้เรียนภาษามากกว่า ส่วนพจนานุกรมภาษาเดียวสำหรับเจ้าของภาษา เช่น Oxford English Dictionary จะบรรจุคำศัพท์จำนวนมากซึ่งรวมถึงศัพท์ยากๆ ที่ไม่ค่อยได้ใช้กัน

พจนานุกรมแบบปัจจุบันถือได้ว่าเริ่มจากพจนานุกรมของ Johnson ในปี ค.ศ.1755 ซึ่งได้รวบรวมคำให้ครอบคลุมทั้งภาษามากกว่าที่จะเก็บเฉพาะคำยากหรือคำยืมจากภาษาต่างประเทศแบบพจนานุกรมแบบเก่า พจนานุกรมของ Johnson นี่เองที่ต่อมาได้ขยายเป็น Oxford English Dictionary ซึ่งตีพิมพ์ฉบับแรกในปีค.ศ.1928 การทำพจนานุกรมยุคแรกนี้ใช้วิธีการรวบรวมข้อความจากหนังสือมาเป็นตัวอย่างการใช้ภาษาที่ดี เป็นข้อมูลในการเลือกคำและให้คำนิยาม นอกจากนี้ยังให้ความสำคัญกับประวัติหรือพัฒนาการ มีการบอกว่าพบคำนั้นใช้ครั้งแรกเมื่อใดในความหมายใด ในกรณีที่คำมีหลายความหมายก็จะจัดลำดับความหมายตามลำดับเวลาการปรากฏ

3. การทำพจนานุกรมแบบใช้คลังข้อมูลภาษา

การทำพจนานุกรมในยุคหลัง เช่น การทำพจนานุกรม Collin Cobuild มีการนำเทคโนโลยีคอมพิวเตอร์เข้ามาช่วย โดยรวบรวมข้อมูลที่เป็นตัวแทนของภาษาอังกฤษจากแหล่งต่างๆ เช่น หนังสือ นิตยสาร การสนทนา แผ่นพับ โฆษณา เป็นต้น การใช้คอมพิวเตอร์ทำให้ได้ข้อมูลที่แต่เดิมไม่สามารถหาได้ คือ ข้อมูลความถี่ของคำ ซึ่งก็มีส่วนช่วยในการตัดสินใจว่าจะบรรจุคำนั้นในพจนานุกรมหรือไม่ แม้ว่าขั้นตอนการให้นิยาม การหาคำพ้องความหมาย ยังเป็นสิ่งที่ผู้ทำพจนานุกรมต้องวิเคราะห์เอง แต่การมีผลที่เป็นคอนคอร์แด๊นซ์ออกมาก็ช่วยให้กระบวนการวิเคราะห์เป็นไปด้วยความสะดวกมากขึ้น

คลังข้อมูลภาษามีบทบาทที่สำคัญต่อการทำพจนานุกรมในปัจจุบัน เพราะเป็นแหล่งข้อมูลที่ช่วยให้ผู้ทำพจนานุกรมได้ศึกษาและเข้าใจความหมายต่างๆของการใช้คำแต่ละคำ นอกจากนี้ การรวบรวมข้อมูลภาษาอย่างต่อเนื่องจะช่วยให้เราสามารถติดตามคำเกิดใหม่และการเปลี่ยนแปลงของความหมายต่างๆ ของคำได้ แต่เนื่องจากการทำพจนานุกรมแต่ละฉบับนั้นมีปัจจัยต่างๆ ที่ต้องคำนึง เช่น กลุ่มผู้ใช้คือใคร มีความรู้ระดับไหน คาดหวังว่าจะได้ข้อมูลอะไรจากพจนานุกรม ใช้ประโยชน์จากพจนานุกรมเพื่อการเขียนอ่านหรือฟังพูด ปัจจัยเหล่านี้จะเป็นตัวกำหนดว่าคลังข้อมูลภาษาที่ควรนำมาใช้เป็นทรัพยากรของการทำพจนานุกรมจะมีลักษณะภาษาแบบใด ขนาดเท่าใด หากเป็นพจนานุกรมทั่วไป คลังข้อมูลภาษาที่ใช้ก็ต้องมีขนาดใหญ่และครอบคลุมข้อมูลภาษากลุ่มประเภทต่างๆ แต่หากเป็นพจนานุกรมเฉพาะทางหรือประมวลศัพท์ คลังข้อมูลภาษาที่ใช้จะมีขนาดที่เล็ก ขนาดของพจนานุกรมหรือจำนวนศัพท์ก็เป็นอีกปัจจัยสำคัญในการกำหนดขนาดของคลังข้อมูลที่ต้องการ Heid (2009) สรุปจากงานที่ผ่านมาว่าหากต้องการทำพจนานุกรมทั่วไปขนาด 50,000-60,000 คำ คลังข้อมูล

ที่ใช้ควรมีขนาดประมาณ 60-100 ล้านคำ แต่หากเป็นพจนานุกรมเฉพาะทางที่มีศัพท์ไม่กี่พันคำ คลังข้อมูลขนาดหนึ่งล้านคำก็มีความเหมาะสมแล้ว

Heid (2009) กล่าวถึงการใช้คลังข้อมูลภาษาในงานทำพจนานุกรมทั้งในระดับภาพรวม (macro) และในระดับย่อย (micro) ในระดับภาพรวม คลังข้อมูลใช้ช่วยคัดเลือกคำที่ควรบรรจุเป็นคำหลักในพจนานุกรมตลอดไปจนถึงสัดส่วนของกลุ่มคำประเภทต่างๆ ตามที่ปรากฏในการใช้จริง ความถี่การปรากฏเป็นข้อมูลสำคัญในการคัดเลือกคำ ในระดับย่อย คลังข้อมูลใช้ช่วยในการแจกแจงความหมายต่างๆ ของคำ ความสัมพันธ์ทางวากยสัมพันธ์กับคำอื่นๆ คำที่มักปรากฏใช้ร่วมกัน บริบทการใช้คำตลอดไปจนถึงทำเนียบภาษาและวิธภาษาที่เหมาะสมกับคำนั้น คำพ้องความหมายและคำตรงข้ามก็เป็นข้อมูลอีกส่วนที่มักให้ไว้ในพจนานุกรม

การทำพจนานุกรมโดยใช้คลังข้อมูลภาษานั้นต้องอาศัยเรื่องทางสถิติและการใช้โปรแกรมคอมพิวเตอร์เป็นเครื่องมือในการทำงาน โปรแกรมคอนคอร์แด๊นซ์ที่แจกแจงรายการคำพร้อมบริบทเป็นโปรแกรมพื้นฐานที่ใช้กันมาตั้งแต่ยุคแรกๆ ของการใช้คอมพิวเตอร์ช่วยทำพจนานุกรม แต่คำจำนวนไม่น้อยที่จะพบตัวอย่างการใช้เป็นจำนวนมากจากคลังข้อมูลซึ่งก็เป็นปัญหาสำหรับผู้ทำพจนานุกรม Kilgarriff (2003) กล่าวถึงเรื่องนี้ไว้ โดยได้ยกตัวอย่างอธิบายว่าในคำที่มีข้อมูลตัวอย่าง 50 ตัวอย่าง ผู้ทำพจนานุกรมสามารถดูข้อมูลและวิเคราะห์ได้โดยง่าย หากเพิ่มเป็น 500 ตัวอย่างก็ยังพอเป็นไปได้แต่ต้องใช้เวลามากขึ้น แต่หากมีตัวอย่างเป็นหลายพันหลายหมื่น ก็เป็นเรื่องยากสำหรับผู้ทำพจนานุกรมที่จะดูรายการตัวอย่างทั้งหมดได้ โปรแกรมคอมพิวเตอร์จึงถูกสร้างขึ้นมาเพื่อช่วยดูข้อมูลจำนวนมากๆนี้ โดยการอาศัยหลักการทางสถิติคอมพิวเตอร์จะสามารถคำนวณหาคำที่มีความสำคัญหรือมีความสัมพันธ์ใกล้ชิด

กับคำหลักที่วิเคราะห์อยู่ได้ ตัวอย่างของโปรแกรมที่ถูกออกแบบมาเพื่อช่วยในการทำพจนานุกรมคือโปรแกรม Word Sketch (Killgariff et al. 2004)

โปรแกรม Word Sketch เป็นโปรแกรมที่ใช้หลักสถิติช่วยหาคำที่มักปรากฏร่วมกับคำหลัก (collocate) แต่มีความสามารถเพิ่มเติมที่แยกคำปรากฏร่วมเป็นกลุ่มๆ ตามหน้าที่ต่างๆ เช่น กลุ่มที่เป็นประธานของคำหลัก กลุ่มที่เป็นกรรมของคำหลัก กลุ่มที่เป็นคำขยาย ฯลฯ เพราะไม่ได้ใช้เพียงการวิเคราะห์รูปคำแต่พิจารณาถึงรูปแบบทางไวยากรณ์ (grammatical pattern) ด้วย โปรแกรมนี้ถูกพัฒนาเพื่อใช้ในการทำพจนานุกรม Macmillion English Dictionary ในปีค.ศ. 2002 ต่อมาจึงได้มีการปรับเพิ่มเติมให้ใช้ภาษาอื่นได้ เช่น จีน ญี่ปุ่น ฝรั่งเศส เยอรมัน ฯลฯ

***pray* (v)** BNC freq= 2455

~ for	680	3.4	~ to	142	1.1	and/or	179	1.7	modifier	338	0.5	object	183	-1.2	subject	1361	0.5
rain	12	19.8	god	32	24.0	hope	20	20.8	silently	15	13.3	god	13	10.5	we	306	12.3
soul	14	19.3	God	22	17.7	hop	13	15.5	together	35	9.3	God	11	9.6	petitioner	7	8.3
-	117	17.3	lord	16	11.4	fast	6	12.2	fervently	4	7.6	prayer	6	7.6	knee	5	6.9
God	11	16.5	saint	4	10.0	pray	16	11.2	aloud	6	7.5	day	9	3.8	congregation	4	6.8
peace	25	16.5	jesus	2	5.4	kneel	5	9.9	earnestly	5	7.3	heaven	2	3.3	i	263	6.2
miracle	8	13.9	emperor	2	5.2	read	9	9.5	inwardly	3	5.5	hook	2	3.3	she	130	5.8
him	26	13.7	Jesus	2	4.5	talk	6	7.4	hard	7	5.3	time	13	3.2	muslim	3	5.7
forgiveness	7	13.4	spirit	2	4.3	sing	4	6.4	daily	3	4.4	night	5	3.1	follower	3	5.0
you	23	13.2	image	2	4.0	watch	4	5.0	only	20	3.8	lord	2	2.7	Jesus	5	4.8
me	24	13.1	wind	2	3.9	live	3	3.9	continually	3	3.7	pardon	2	2.7	jew	3	4.5
deliverance	6	13.0	him	6	3.3	work	5	3.5	regularly	5	3.5	soul	2	2.4	church	7	4.5
them	23	12.2				wish	2	3.4	often	10	3.3	silence	3	2.4	fellowship	2	4.0
church	12	11.7				believe	2	2.9	ever	9	3.0				Singh	2	3.7
guidance	8	11.6				learn	2	2.8	secretly	2	2.7				Family	6	3.6
us	16	11.6				tell	2	2.3	quietly	3	2.4						
chance	5	10.3							still	11	2.3						

รูป 5.1 ตัวอย่างคำปรากฏร่วมของคำว่า *pray* แยกตามประเภทต่างๆ

นอกจากข้อมูลคำปรากฏร่วม กลุ่มคำหรือวลีที่เป็นเสมือนหนึ่งหน่วยภาษา (lexical unit) ก็เป็นอีกข้อมูลหนึ่งที่สำคัญโดยเฉพาะสำหรับผู้ที่ไม่ใช่เจ้าของภาษา เพราะสิ่งที่ดูเหมือนเป็นอะไรที่น่าจะเห็นได้ชัดเจน เช่น การใช้คำว่า *day* มักจะพบเป็นวลีอย่าง *one day*, *the other day*, *some day* แต่หากไม่มี

คลังข้อมูลภาษาให้ค้น ข้อเท็จจริงเหล่านี้ก็ยากที่จะนึกคิดขึ้นมาเองได้ (Summers 2005)

4. อรรถศาสตร์คำศัพท์

เป็นการศึกษาความสัมพันธ์ทางความหมายระหว่างคำในภาษา เกิดขึ้นในยุคที่ภาษาศาสตร์สนใจศึกษาหาระบบของภาษา การศึกษาความหมายจึงเปลี่ยนจากการศึกษาความเป็นมา การเปลี่ยนแปลงของความหมาย มาเป็นการมองหาระบบความหมาย ความสัมพันธ์ทางความหมายระหว่างคำนั้นมีหลากหลาย เช่น ความสัมพันธ์แบบจ่ากลุ่มลูกกลุ่ม ความสัมพันธ์แบบองค์รวมองค์ประกอบ ความสัมพันธ์แบบคำพ้องรูปหรือเสียง ความสัมพันธ์แบบคำพ้องความหมาย ความสัมพันธ์แบบคำตรงข้าม เป็นต้น ความสัมพันธ์ระหว่างคำที่สำคัญมีดังนี้

ความสัมพันธ์แบบคำพ้อง (homonymy) คือความสัมพันธ์ที่คำมีรูปหรือเสียงเหมือนกันแต่มีความหมายแตกต่างกัน เรียก homonym เช่น คำว่า *bank* ที่หมายถึง ธนาคาร หรือ ริมฝั่งน้ำ ก็ได้ กรณีเช่นนี้แสดงว่าคำนี้น่าจะเป็น 2 หน่วยศัพท์ (lexeme) ที่แยกจากกันแต่บังเอิญมีรูปเหมือนกัน *bank1, bank2* ในกรณีที่คำที่มีเสียงเหมือนกันแต่มีความหมายแตกต่าง เช่น *would-wood, be-bee* ก็จัดเป็นคำพ้องได้เรียกว่าคำพ้องเสียง (homophone) ส่วนกรณีที่คำมีรูปเหมือนกันแต่ออกเสียงต่างกัน เช่น bass ที่หมายถึง ปลา จะออกเสียงต่างจากที่หมายถึงเครื่องดนตรี จึงจัดเป็นคำพ้องที่เรียกว่าคำพ้องรูป (homograph)

ความสัมพันธ์แบบหลายความหมาย (polysemy) เป็นกรณีที่คำหนึ่งมีความหมายได้หลายอย่าง แต่ทั้งหมดก็เป็นความหมายที่เกี่ยวข้องกัน จึงไม่ใช่กรณีเดียวกับคำพ้อง ความต่างนี้แสดงในพจนานุกรมด้วย โดยคำพ้องรูปจะถูก

จัดเก็บเป็นคนละหน่วยศัพท์ในพจนานุกรม ในขณะที่คำหลายความหมายจัดเป็นหน่วยศัพท์เดียว แต่ให้อธิบายความหมายต่างๆ ภายใต้คำศัพท์นั้น

ความสัมพันธ์แบบพ้องความหมาย (synonymy) เป็นความสัมพันธ์ระหว่างคำที่จัดได้ว่ามีความหมายเหมือนกัน สามารถใช้แทนกันได้ ปกติคำที่พ้องความหมายแท้จริงที่สามารถใช้แทนกันได้ในทุกบริบททุกสถานการณ์นั้นเป็นไปได้ยาก คำพ้องความหมายจึงใช้ในความหมายที่อ่อนลงคือหมายถึงคำที่สามารถใช้แทนกันได้ในบางบริบทมากกว่า ในพจนานุกรมจะจัดเก็บเป็นคนละหน่วยศัพท์แต่มักมีการบอกว่าคำที่พ้องความหมายของหน่วยศัพท์นั้นคืออะไร

ความสัมพันธ์แบบจ่ากลุ่ม (hypenymy) และแบบลูกกลุ่ม (hoponymy) เป็นความสัมพันธ์ที่หน่วยศัพท์หนึ่งเป็นประเภทย่อย (subclass) ของอีกหน่วยศัพท์ที่เป็นประเภทใหญ่กว่าหน่วยศัพท์ใหญ่กว่าเรียกคำจ่ากลุ่ม (hypernym) หน่วยศัพท์ย่อยกว่าเรียกคำลูกกลุ่ม (hyponym) ถ้า x เป็นคำลูกกลุ่มของ y ประโยคที่มี x เป็นจริง ประโยคที่มี y ก็จะเป็นจริงด้วย เช่น ถ้า x = car y = vehicle ถ้าประโยค *This is a car* เป็นจริง *This is a vehicle* ก็จะเป็นจริงด้วย ความสัมพันธ์แบบจ่ากลุ่มลูกกลุ่มนี้ช่วยให้เราสามารถจัดจำพวก (taxonomy) ของมโนทัศน์ต่างๆได้ เป็นโครงสร้างมโนทัศน์ที่มีลำดับชั้น (hierarchy) อย่างไรก็ตาม คำลูกกลุ่มของคำจ่ากลุ่มเดียวกันอาจจะไม่สามารถรวมเป็นกลุ่มประเภทเดียวกันได้ เช่น *hound, mutt, puppy* ต่างก็เป็นคำลูกกลุ่มของ *dog* แต่จัดรวมกันไม่ได้เพราะคำเหล่านี้เป็นการแยกย่อยด้วยเกณฑ์ที่ต่างกัน *hound* เป็นการแยกย่อยตามหน้าที่เป็นสุนัขล่าเนื้อ *mutt* เป็นการแยกตามสายพันธุ์เป็นสุนัขพันธ์ทาง *puppy* เป็นการแยกตามอายุเป็นลูกสุนัข

ความสัมพันธ์แบบองค์รวม (holonymy) กับแบบองค์ประกอบ (meronymy) เป็นความสัมพันธ์ที่บอกว่าหน่วยศัพท์หนึ่งเป็นองค์ประกอบของอีก

หน่วยศัพท์หนึ่ง เช่น ประตูเป็นส่วนหนึ่งของบ้าน ลูกบิดเป็นส่วนหนึ่งของประตู เป็นต้น ความสัมพันธ์แบบนี้ไม่สามารถส่งข้ามขั้นได้เสมอหรือมีข้อจำกัดการส่งข้าม (transitivity) เช่น ถ้าบอกว่าลูกบิดเป็นส่วนหนึ่งของบ้าน ก็จะแปลก

ความสัมพันธ์ระหว่างคำต่างๆ ที่กล่าวมานั้นมีเพียงบางส่วนเท่านั้นที่บรรจุลงในพจนานุกรม ความสัมพันธ์แบบหลายความหมายเป็นความสัมพันธ์พื้นฐานของคำที่พบในพจนานุกรมที่มีการแยกอธิบายแต่ละความหมายในคำหลัก ความสัมพันธ์แบบคำพ้องความหมาย ความหมายตรงข้าม มีบรรจุอยู่ในพจนานุกรมบางเล่ม พจนานุกรมที่ให้ความสัมพันธ์กับคำพ้องความหมาย คำตรงข้าม คือพจนานุกรมที่เรียกว่าอรรถาภิธาน (thesaurus) นอกจากอรรถาภิธานแล้ว ยังมีพจนานุกรมอีกประเภทที่ทำขึ้นมาเพื่อใช้กับคอมพิวเตอร์ในงานวิจัยที่เกี่ยวข้องกับทางการประมวลผลภาษาคือ เครือข่ายคำ (wordnet)

5. เครือข่ายคำ

WordNet หรือเครือข่ายคำเป็นคลังคำที่สร้างขึ้นเพื่อสะท้อนความสัมพันธ์ทางความหมายระหว่างคำต่างๆ เริ่มต้นพัฒนาขึ้นภายใต้การนำของ George A. Miller มหาวิทยาลัย Princton ตั้งแต่ปีค.ศ.1985 ซึ่งตอนนั้นมีลักษณะเป็นการทดลองเล็กๆ จนกระทั่งได้มีการพัฒนาเป็นเครือข่ายคำที่มีขนาดใหญ่และใช้งานได้จริงและได้ออกรุ่นแรกในปี ค.ศ.1991 จนถึงปัจจุบันมีการพัฒนาเป็นรุ่น WordNet 3.0

Fellbaum (1998) อธิบายว่าใน WordNet มีการจัดกลุ่มคำที่มีความหมายเดียวกันหรือใช้แทนมโนทัศน์เดียวกันเป็นกลุ่มเรียกว่า synset จากนั้นจึงเชื่อมโยงแต่ละมโนทัศน์ด้วยความสัมพันธ์ทางความหมายแบบต่างๆ WordNet มีลักษณะผสมผสานระหว่างพจนานุกรมกับอรรถาภิธาน เพราะมีการให้นิยามความหมายของแต่ละมโนทัศน์มีการให้ตัวอย่างการใช้ ในขณะเดียวกันก็

จัดกลุ่มคำที่มีความหมายเดียวกัน แสดงความสัมพันธ์ทางความหมายแบบระดับชั้น แต่ใน WordNet ก็มีข้อมูลความสัมพันธ์ทางความหมายแบบอื่นๆ ด้วย ข้อมูลที่ใส่ใน synset ก็อาจเป็นวลีก็ได้หากมโนทัศน์นั้นไม่มีคำเดี่ยวที่สื่อมโนทัศน์นั้นได้ เช่น *offender* และ *libertine* เป็นมโนทัศน์ย่อยของ *bad person* เป็นต้น ข้อมูลที่เก็บใน WordNet มีเฉพาะคำที่เป็นหมวดคำเปิดสี่ประเภท คือ nouns, verbs, adjectives และ adverbs ซึ่งจะเก็บแยกจากกัน

WordNet รุ่น 3.0 มีจำนวนข้อมูลตามตาราง คือ คอลัมน์ที่สองบอกจำนวนรูปคำที่ปรากฏในแต่ละหมวด คอลัมน์สามบอกจำนวน synset คอลัมน์สุดท้ายบอกจำนวนการเชื่อมโยงระหว่างรูปคำกับ synset เช่น คำนามมีรูปคำแตกต่างกัน 117,798 คำ จัดเป็น synset 82,115 กลุ่ม ซึ่งเชื่อมโยงระหว่างรูปคำกับ synset จำนวน 146,312 การเชื่อมโยง เป็นต้น

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

ตาราง 5.1 แจกแจงข้อมูลใน WordNet 3.0

ตัวอย่างข้างล่างแสดงข้อมูลที่ได้จากค้นคำว่า *plant* ใน WordNet ซึ่งจะได้ว่าคำนี้มี 4 ความหมายหรือมโนทัศน์ ในแต่ละมโนทัศน์จะมีคำพ้องความหมายหรือที่อยู่ในมโนทัศน์นั้น เช่น ในความหมายที่หนึ่งมีคำว่า *plant, works,* และ *industrial plant* ส่วนนิยามหรือคำอธิบายความหมายแสดงในวงเล็บพร้อมทั้งตัวอย่างการใช้

Sense 1 : plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")

Sense 2 : plant, flora, plant life -- ((botany) a living organism lacking the power of locomotion)

Sense 3 : plant -- (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)

Sense 4 : plant -- (something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant")

หากดูข้อมูลที่เก็บจริงในไฟล์ noun.index จะพบว่า *plant* เป็นคำนาม (n) มี 4 ความหมาย ได้แก่ มโนทัศน์รหัส 03956922 รหัส 00017222 รหัส 10438470 และรหัส 05906080 และเมื่อดูข้อมูลในไฟล์ noun.data ซึ่งแสดงรายการคำที่อยู่ในมโนทัศน์ต่างๆ เช่น plant ในมโนทัศน์ 03956922 มีคำที่จัดอยู่ในมโนทัศน์นี้คือ {*plant, works, industrial_plant*} พร้อมทั้งบอกความสัมพันธ์ทางความหมายกับมโนทัศน์อื่นๆ เช่น มีมโนมัศน์จ่ากลุ่ม (@) เป็น 02914991 (*building_complex* และ *complex*) มีมโนทัศน์ลูกลุ่ม (~) 11 มโนทัศน์

ข้อมูลจากไฟล์ noun.index

plant n 4 7 @ ~ #m %p + ; - 4 2 03956922 00017222 10438470 05906080

ข้อมูลจากไฟล์ noun.data

03956922 06 n 03 plant 1 works 2 industrial_plant 0 013 @ 02914991 n 0000 ~ 02878107 n 0000 ~ 02897237 n 0000 ~ 03212811 n 0000 ~ 03316406 n 0000 -c 03425956 n 0000 ~ 03772269 n 0000 ~

03871860 n 0000 ~ 04065909 n 0000 ~ 04068441 n 0000 ~ 04131811 n 0000 ~ 04178668 n 0000 ~ 04246156 n 0000 | buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles"

ความสัมพันธ์ที่มีอยู่ในเครือข่ายคำมีความสัมพันธ์สองแบบ แบบแรกคืออรรถสัมพันธ์ (semantic relation) ซึ่งเป็นความสัมพันธ์เชื่อมโยงระหว่าง synset เช่น ความสัมพันธ์แบบจ่ากลุ่ม (hyperymy) ลูกกลุ่ม (hyponymy) ความสัมพันธ์แบบองค์ประกอบ (meronymy) องค์รวม (holonymy) และความสัมพันธ์แบบ entailment ความสัมพันธ์ประเภทที่สองคือ ศัพทสัมพันธ์ (lexical relation) ที่เชื่อมโยงระหว่างรูปศัพท์ ได้แก่ ความสัมพันธ์แบบคำพ้องความหมาย (synonymy) และความสัมพันธ์แบบคำตรงข้าม (antonymy) ความสัมพันธ์แบบคำพ้องความหมายเป็นความสัมพันธ์ระหว่างรูปศัพท์ในตัวอยู่แล้วเพราะเป็นการจัดกลุ่มคำซึ่งมีความหมายเหมือนกัน ส่วนความสัมพันธ์แบบคำตรงข้ามก็เป็นความสัมพันธ์ระหว่างรูปศัพท์ เพราะเราจะเห็นว่าคำที่มีความหมายเหมือนหรือใกล้เคียงกันไม่จำเป็นต้องมีคำตรงข้ามเหมือนกัน เช่น *heavy* มีความหมายเหมือนกับ *weighty* แต่ *light* เป็นคำตรงข้ามของ *heavy* ไม่ได้เป็นคำตรงข้ามของ *weighty* ด้วย ดังนั้น จึงมีความต่างระหว่างการมีความหมายตรงข้ามกับการมีคำตรงข้าม

WordNet ไม่ได้มีความสัมพันธ์ทางความหมายทุกอย่างไว้ เช่น ไม่ได้ให้ความสัมพันธ์ระหว่างมโนทัศน์ที่เป็นคำนามกับมโนทัศน์ที่เป็นกริยา ไม่ได้แยกย่อยความสัมพันธ์แบบจ่ากลุ่มลูกกลุ่ม ซึ่งอาจมองเป็นแบบ is-a-kind-of คือมองแบบจัดจำพวกหรือจัดตามรูป (taxonomic หรือ formal) หรือ is-used-as-a-kind-of คือจัดตามหน้าที่ (functional หรือ telic) ตัวอย่างเช่น a chicken is a

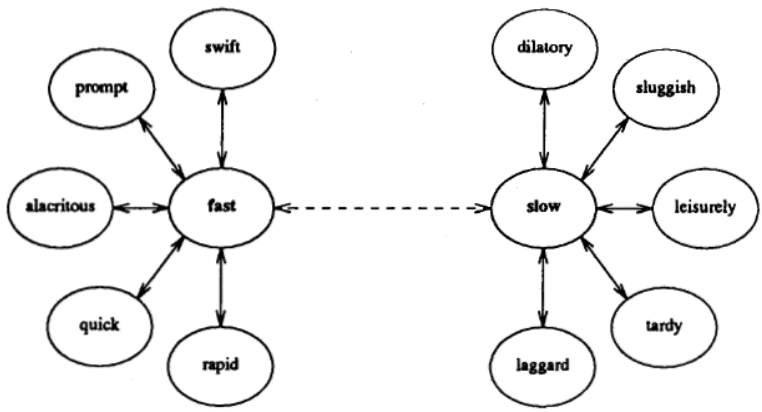
bird เป็นการจัดจำพวก แต่ a chicken is a food เป็นการจัดตามหน้าที่ สาเหตุที่ความสัมพันธ์จ่ากลุ่มลูกกลุ่มไม่ได้แยกประเภทย่อยลักษณะนี้ เป็นเพราะ WordNet ได้สร้างขึ้นมาแล้วเป็นจำนวนมากก่อนที่ผู้สร้างจะตระหนักถึงข้อแตกต่างนี้ จึงจำต้องปล่อยให้เป็นไป (Fellbaum 1998:35) ผู้ใช้จึงต้องรู้เองว่ามีความต่างเช่นนี้อยู่ ความสัมพันธ์แบบจ่ากลุ่ม เช่น {robin, redbreast}@->{bird} @->{animal, animate_being}@->{organism, life_form, living_thing} ทำให้เกิดเป็นโครงสร้างลำดับชั้นของมโนทัศน์เหมือน ontology ซึ่งเป็นโครงสร้างความรู้ (knowledge structure) ที่มโนทัศน์ล่างสามารถรับทอดคุณสมบัติ (inherit property) จากโหนดบนมาได้ นอกจากนี้ WordNet ก็ไม่ได้ระบุว่ามโนทัศน์ไหนเป็นมโนทัศน์พื้นฐาน (basic concept) แบบที่ใช้กันในทางภาษาศาสตร์จิตวิทยา

ในความสัมพันธ์แบบองค์รวมองค์ประกอบ ซึ่งทดสอบได้จาก x is a part of y ความสัมพันธ์แบบนี้ใน WordNet แยกประเภทย่อยเป็นสามชนิด คือ 1) เป็นส่วนประกอบ (is a component part of) เช่น *branch* เป็นส่วนประกอบหนึ่งของ *tree* 2) เป็นสมาชิก (is a member of) เช่น *tree* เป็นสมาชิกของ *forest* 3) ทำขึ้นจาก (is made from) *airplane* ทำจาก *aluminium* ซึ่งความสัมพันธ์ที่พบมากสุดใน WordNet คือ การเป็นส่วนประกอบ (Fellbaum 1998:39)

คำคุณศัพท์ใน WordNet มีทั้งที่เป็นคุณศัพท์พรรณา (descriptive adjective) และคุณศัพท์สัมพันธ์ (relational adjective) คุณศัพท์พรรณาบอกถึงคุณลักษณะ (attribute) ของคำนาม เช่น *heavy* หรือ *light* เป็นค่า (value) ของคุณลักษณะที่เป็นน้ำหนักของสิ่งต่างๆ ความสัมพันธ์ที่พบในหมวดคำคุณศัพท์คือความสัมพันธ์แบบคำตรงข้ามซึ่งเป็นความสัมพันธ์ระหว่างคำไม่ใช่มโนทัศน์ดังที่กล่าวมาแล้ว คำที่มีความหมายเหมือนหรือใกล้กันไม่ได้มีคำตรงข้ามคำเดียวกัน เช่น คำ *heavy, weighty, ponderous* ไม่ได้มีคำตรงข้ามเป็น *light,*

weightless, *airy* แต่เป็นว่า *heavy* ตรงข้ามกับ *light*, *weighty* ตรงข้ามกับ *weightless*, และ *ponderous* ตรงข้ามกับ *airy*

ในกลุ่มคำคุณศัพท์จึงมีความสัมพันธ์แบบคล้ายคลึง (similar) ที่จัดกลุ่มคำคุณศัพท์ที่มีความหมายใกล้เคียงกันแต่ไม่ได้มีคำตรงข้ามเหมือนกัน ดังตัวอย่างในรูปข้างล่างนี้ที่ *slow* เป็นคำตรงข้ามกับ *fast* แต่ไม่ได้เป็นคำตรงข้ามกับ *quick* หรือ *rapid* คำอย่าง *quick*, *rapid* จึงเป็นคำคล้ายคลึงกับ *fast* (ต่างจากความสัมพันธ์แบบคำพ้องความหมายที่จะจัดเป็น synset หรือมโนทัศน์เดียวกัน) และเป็นคำตรงข้ามทางอ้อม (indirect antonym) ของ *slow* (คือมีความหมายที่ตรงข้ามกับ *slow* แต่ไม่ได้เป็นคำตรงข้ามของ *slow*)



รูป 5.2 แสดงคำตรงข้ามทางอ้อมคำคล้ายคลึง

(Fellbaum 1998:51)

WordNet เชื่อมโยงความสัมพันธ์ระหว่างคุณศัพท์พรรณาและคำนามที่เกี่ยวข้องกันโดยการสร้างลิ้งก์ระหว่างคำที่เรียกว่าคุณลักษณะ (attribute) เช่น *depth* มีคุณลักษณะที่เกี่ยวข้องคือ *deep* กับ *shallow* *beautiful* เป็นคุณลักษณะของ *beauty* เป็นต้น แต่ใน WordNet ก็ไม่มีการบอกข้อมูลว่า

คำคุณศัพท์ไหนเป็นคำสามัญกว่า (unmarked) เช่น คำว่า *deep* กับ *shallow* แม้จะเป็นคำตรงข้ามกัน แต่เราใช้ *The pool is five feet deep* ได้ แต่ใช้ **The pool is five feet shallow* ไม่ได้ คำคุณศัพท์ *deep* จึงเป็นคำที่สามัญกว่า

คำคุณศัพท์สัมพันธ์ (relational adjective) เป็นคำซึ่งมีความหมายเกี่ยวข้องกับคำนาม เช่น *musical* เกี่ยวข้องกับ *music* *dental* เกี่ยวข้องกับ *tooth* คำคุณศัพท์สัมพันธ์นี้เมื่อใช้ขยายนามจะไม่ได้ใช้เพื่อบอกคุณลักษณะ (attribute) ของคำนามนั้น แต่บอกความสัมพันธ์ลักษณะอื่น ซึ่งแตกต่างกันไปเมื่อใช้กับคำนามต่างกันได้ เช่น *agricultural equipment* บอกว่าเครื่องมือนั้นใช้เพื่อการเกษตร แต่ *agricultural college* บอกว่าวิทยาลัยนั้นสอนการเกษตร คำคุณศัพท์สัมพันธ์ต่างจากคำคุณศัพท์พรรณาที่ไม่มีคำตรงข้าม ไม่มีการให้ระดับ (gradable) เช่น **the extremely atomic bomb* (Fellbaum 1998:59-60)

คำกริยาวิเศษณ์ (adverb) เนื่องจากเกี่ยวข้องกับคำคุณศัพท์ จึงมีลิ้งก์ derived-from โยงกับคำคุณศัพท์ เช่น *beautifully* ลิ้งก์กับ *beautiful* คำกริยาวิเศษณ์ไม่มีการจัดลำดับชั้นเหมือนคำนาม คำกริยา ไม่มีการจัดเป็นกลุ่ม (cluster) เหมือนกับคำคุณศัพท์ แต่อาจมีคำตรงข้ามได้ซึ่งก็ได้มาจากคำคุณศัพท์ เช่น *specifically* มีคำตรงข้ามคือ *generally* เหมือนกับที่ *specific* เป็นคำตรงข้ามของ *general*

คำกริยาใน WordNet ก็มีการจัดกลุ่มตามวงความหมาย (semantic domain) เหมือนกับการจัดกลุ่มกริยาโดยทั่วไป ซึ่งมักแยกเป็นกริยาบอกการกระทำ (action verb) กริยาบอกสภาพ (state verb) ใน WordNet มีการแบ่งประเภทย่อยลงไปอีก ได้แก่ กริยาบอกการเคลื่อนที่ (motion) การรับรู้ (perception) การสัมผัส (contact) การสื่อสาร (communication) การแข่งขัน (competition) การเปลี่ยนแปลง (change) ความรู้สึก (cognition) การบริโภค

(consumption) การสร้าง (creation) อารมณ์ (emotion) การครอบครอง (possession) การดูแลร่างกาย (bodily care) การทำงานร่างกาย (bodily function) พฤติกรรมในสังคม (social behavior and interaction) และยังมีกลุ่มที่บอกสภาพ เช่น *be, resemble, belong* กลุ่มนี้รวมถึงคำกริยาช่วยและ control verb อย่าง *want, fail, prevent, succeed* ตลอดจน aspectual verb อย่าง *begin* รวมทั้งสิ้นมี 15 กลุ่ม อย่างไรก็ตาม การแบ่งก็ไม่ได้เป็นอะไรที่ขาดจากกันชัดเจน คำกริยาหนึ่งอาจจัดอยู่ในวงความหมายมากกว่าหนึ่งวงได้ เช่น *wonder* เป็นได้ทั้ง cognition และ communication และในแต่ละวงความหมายก็ไม่ได้มีมโนทัศน์บนสุดเพียงมโนทัศน์เดียว อาจมีมากกว่าหนึ่งมโนทัศน์ที่ต่างก็แยกย่อยต่อไปอีกได้ (Fellbaum 1998:70-72) อย่างไรก็ตาม การจัดวงความหมายเป็นไปเพื่อช่วยผู้จัดทำ WordNet แต่ไม่ได้เกี่ยวข้องโดยตรงสำหรับผู้ใช้เพราะผู้ใช้เรียกดูข้อมูลคำกริยาแต่ละตัวก็จะเห็นข้อมูลความสัมพันธ์ที่เกี่ยวกับคำกริยานั้น

คำกริยามีความสัมพันธ์แบบคำตรงข้าม เช่น *start-stop, remember-forget, ask-tell* เป็นต้น มีความสัมพันธ์แบบ entailment ซึ่งเป็นความสัมพันธ์แบบตรรกะ หากกริยา V1 entail V2 ถ้าประโยคที่ใครทำ V1 เป็นจริง ประโยคที่ใครทำ V2 ก็จะเป็นจริงด้วย แต่ไม่เป็นจริงในทิศทางตรงข้าม ตัวอย่างเช่น *snore* จะ entail *sleep* เพราะทุกครั้งที่มีการกรนคนๆ นั้นจะต้องหลับอยู่ แต่คนที่หลับไม่จำเป็นต้องกรน ถ้าคำกริยาสองคำ entail ซึ่งกันและกันแสดงว่าคำกริยาทั้งสองเป็นคำพ้องความหมายหรืออยู่ใน synset เดียวกัน

ความสัมพันธ์แบบ troponymy เป็น entailment ประเภทหนึ่ง หากคำกริยา V1 เป็น troponym ของ V2 V1 จะ entail V2 ด้วย การทำกริยา V1 คือการทำกริยา V2 ด้วยลักษณะบางประการ (to V1 is to V2 in some particular

manner) ตัวอย่างเช่น *march* เป็น troponym ของ *walk* (เป็นลักษณะการเดินประเภทหนึ่ง) การ *march* จึง entail การ *walk* ด้วย troponymy เกี่ยวข้องกับเรื่องของเวลา คำกริยาทั้งสองจะเกิดไปพร้อมกัน (temporally coextensive) คือแต่ละก้าวที่เป็นการ *march* นั้นก็จะเป็นการ *walk* ด้วย ต่างจาก *snore* ที่ entail *sleep* แต่ไม่ได้เป็น troponym ของ *sleep* ความสัมพันธ์แบบ troponymy ทำให้จัดจำพวกกริยาเป็นลำดับชั้นได้ในลักษณะเดียวกับที่จัดคำนามเป็นคำจ่ากลุ่มลูกกลุ่ม ซึ่งลำดับชั้นในคำกริยามีจำนวนระดับที่น้อยกว่าของคำนาม

ความสัมพันธ์ cause ก็เป็น entailment อีกประเภทหนึ่ง ถ้าหาก V1 เป็นเหตุให้เกิด V2 V1 ก็จะ entail V2 เช่น การ *expel* ทำให้เกิดการ *leave* การ *feed* ทำให้เกิดการ *eat* ซึ่งทิศทางความสัมพันธ์จะเป็นทิศทางเดียว เพราะการ *eat* ไม่ได้ entail ว่ามีการ *feed* ความสัมพันธ์ลักษณะนี้ไม่มีเรื่องของการเกิดในเวลาช่วงเดียวกันเหมือนแบบ troponymy

ในคำกริยามีความสัมพันธ์แบบคำตรงข้าม เช่น *give* ตรงข้ามกับ *take*, *buy* ตรงข้ามกับ *sell*, *live* ตรงข้ามกับ *die*, *wake* ตรงข้ามกับ *sleep* คำตรงข้ามนี้บางคำมีลักษณะ entailment ด้วย เช่น *fail* ซึ่งตรงข้ามกับ *succeed* ทั้งสองคำจะ entail *try*, *forget* จะ entail *know* ลักษณะคำตรงข้ามประเภทนี้เรียก backward presupposition ซึ่งอยู่ในกลุ่มที่ไม่มีเรื่องของการเกิดในเวลาช่วงเดียวกันมาเกี่ยวข้อง

ความสัมพันธ์ของกริยาใน WordNet ไม่ได้ครอบคลุมความสัมพันธ์ที่มีระหว่างกริยาและคำนามเหมือนที่ Fillmore และ Atkins ทำในโครงการพจนานุกรม Framenet ที่อธิบายในรูปของกรอบความหมาย (frame semantics) ว่ากริยานั้นมีองค์ประกอบอะไรที่มาเกี่ยวข้องด้วย ในโครงการ WordNet บอกเพียงกรอบวากยสัมพันธ์ (syntactic frame) ของกริยานั้นเป็นอย่างไร

6. ศัพทวิทยาและการทำประมวลศัพท์

Terminology หรือศัพทวิทยาเป็นศาสตร์ที่มองได้ว่าคล้ายกับพจนวิทยา หรือ lexicology และการทำประมวลศัพท์ (terminology หรือ terminography) ก็คล้ายกับการทำพจนานุกรม (lexicography) สิ่งที่แตกต่างกันคือขอบเขตเนื้อหา พจนานุกรมจะครอบคลุมภาษาโดยรวมในขณะที่ประมวลศัพท์จะครอบคลุมศัพท์เฉพาะสาขา

Wüster เป็นคนแรกที่แยกความต่างของการทำประมวลศัพท์ออกจากการทำพจนานุกรมอย่างชัดเจนในงานวิทยานิพนธ์ปริญญาเอกของเขาในช่วงทศวรรษ 1930 ความจำเป็นเรื่องของการทำประมวลศัพท์เริ่มเห็นชัดเจนมากขึ้นตั้งแต่ศตวรรษที่ 18 ในยุคที่วิทยาศาสตร์เริ่มขยายตัวทำให้มีการค้นพบความรู้ใหม่ๆ มากมาย ความรู้เหล่านั้นจำเป็นต้องมีคำศัพท์ที่จะใช้สื่อความหมายใหม่ๆ ที่เกิดขึ้น (Cabré and Sager 1999)

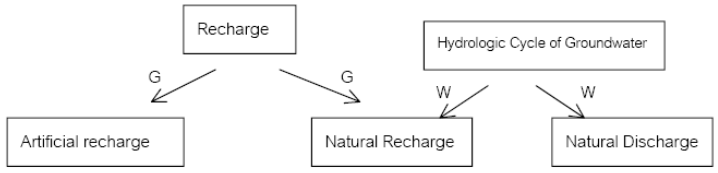
นักวิชาการที่ทำงานด้านประมวลศัพท์มองว่าการทำประมวลศัพท์นั้นแตกต่างจากการทำพจนานุกรมอยู่หลายประการ ในขณะที่พจนานุกรมเป็นการมองจากรูปคำเพื่ออธิบายความหมายของคำนั้น ประมวลศัพท์เป็นเรื่องของการหารูปศัพท์ที่แทนความหมายหรือมโนทัศน์ที่มีอยู่ในสาขาวิชานั้น ศัพท์มีคุณสมบัติที่แตกต่างจากคำทั่วไปในหลายๆมิติ ตั้งแต่ในเรื่องผู้ใช้ เนื้อหาที่คำและศัพท์สื่อ ลักษณะปริจเฉทที่ใช้ ดังนี้

- วิธีการสร้างศัพท์ (term formation) มีการใช้ที่แตกต่างจากวิธีการสร้างคำ (word formation) ประมวลศัพท์จะมีศัพท์ที่มีความซับซ้อนทางโครงสร้างมากกว่า ได้แก่ เป็นคำประกอบ (multi-word เช่น front-wheel-drive vehicle) เป็น set phrase (เช่น black and white) เป็น phrasal term (เช่น bridges with pin-joined members)

- ศัพท์ในประมวลศัพท์ส่วนมากเป็นคำนาม คำกริยา ส่วนคำในพจนานุกรมครอบคลุมหมวดคำทุกหมวด
- การทำประมวลศัพท์ไม่สนใจวิเคราะห์วจีวิภาคหรือแยกองค์ประกอบของศัพท์เหมือนการทำพจนานุกรม
- ประมวลศัพท์ให้ความสำคัญกับมโนทัศน์มากกว่าและถือว่าเป็นสิ่งที่มีมาก่อนและแยกจากรูปศัพท์ได้ ในขณะที่คำในพจนานุกรมถูกมองว่ามีรูปและมโนทัศน์ที่ไม่สามารถแยกจากกันได้ซึ่งเป็นมุมมองแบบนักภาษาศาสตร์
- คำในพจนานุกรมมักมีหลายความหมายได้ (polysemy) แต่ละความหมายจะแสดงในรายการคำเดียวกัน ศัพท์ในประมวลศัพท์โดยปกติมีความหมายเดียวหรือแทนมโนทัศน์เดียว หากมีได้หลายความหมายจะมองเป็นคำพ้องรูปและแสดงเป็นคนละรายการศัพท์
- การทำพจนานุกรมเริ่มด้วยการรวบรวมคำต่างๆที่มีแล้วมาอธิบายการใช้ต่างๆ ส่วนการทำประมวลศัพท์เริ่มจากรายการมโนทัศน์ หาโครงสร้างความสัมพันธ์ระหว่างมโนทัศน์ และหาศัพท์ที่แทนแต่ละมโนทัศน์
- การทำพจนานุกรมเป็นไปเพื่ออธิบายความรู้เรื่องคำในภาษา ส่วนการทำประมวลศัพท์เป็นไปเพื่อสร้างมาตรฐานการใช้ศัพท์ให้เป็นมาตรฐานเดียวกัน
- ผู้ใช้ศัพท์จะเป็นผู้เชี่ยวชาญในสาขา ส่วนผู้ใช้คำเป็นบุคคลทั่วไป

การทำประมวลศัพท์เริ่มต้นจากการรวบรวมคลังข้อมูลในสาขาเพื่อเป็นตัวแทนของสาขาวิชานั้น จากนั้นจึงดึงศัพท์เฉพาะในสาขานั้นออกมา วิเคราะห์ความหมายเพื่อหาว่ามโนทัศน์ต่างๆ มีความสัมพันธ์กันอย่างไร มโนทัศน์สัมพันธ์ที่สร้างขึ้นได้เป็นภาพแทนองค์ความรู้ในสาขาวิชานั้น (ดูตัวอย่างรูป 5.3) การวิเคราะห์ต่าง ๆ นี้เป็นการวิเคราะห์โดยอิงคลังข้อมูลภาษาที่สร้างขึ้น ตัวอย่างการ

ใช้และความสัมพันธ์ระหว่างศัพท์ต่างๆ เป็นสิ่งที่สามารถสกัดได้จากข้อมูลโดยตรง เมื่อได้ภาพการวิเคราะห์ที่ชัดเจนแล้วจึงเขียนประมวลศัพท์แต่ละศัพท์ออกมา



*No4: **B**asic **E**lements*

G = Generic – Specific

W = Whole - Part

รูป 5.3 ตัวอย่างแสดงการวิเคราะห์มโนทัศน์สัมพันธ์ระหว่างศัพท์ต่างๆ

การเขียนนิยามศัพท์มีวิธีการหลากหลาย แต่โดยหลักการคือการเขียนนิยามเพื่อให้ผู้อ่านระบุเห็นภาพมโนทัศน์นั้นภายในระบบมโนทัศน์ทั้งหมดในสาขานั้นได้ โดยทั่วไปจะใช้วิธีการแบบ analytic คืออ้างถึงมโนทัศน์ใหญ่ แล้วให้ข้อมูลเพิ่มเติมเพื่อให้แยกมโนทัศน์นั้นจากมโนทัศน์ข้างเคียงได้ (class + differences) ดังตัวอย่างข้างล่าง

printer = A computer peripheral that outputs data to hardcopy.

laser printer = A nonimpact printer that uses a low-power laser to produce image-forming charges on the photoconductive surface of a drum.

การให้นิยามลักษณะนี้เหมาะกับการทำประมวลศัพท์สำหรับผู้ที่ใช้ที่อยู่ในสาขานั้น เพราะเป็นนิยามที่กระชับและช่วยให้ผู้ใช้ระบุตำแหน่งของมโนทัศน์ในระบบมโนทัศน์ที่ผู้ใช้ในสาขานั้นมีอยู่แล้วได้ แต่หากเป็นประมวลศัพท์ที่ทำเพื่อผู้ใช้ที่ไม่มีความรู้ในสาขามากพอ ก็อาจต้องให้ข้อมูลเพิ่มเติมมากขึ้น

7. บทสรุป

คลังข้อมูลภาษาเป็นทรัพยากรที่จำเป็นและสำคัญสำหรับการทำพจนานุกรมและประมวลศัพท์ในปัจจุบัน ทำให้ได้ข้อมูลที่ครบถ้วนและทันสมัย ขนาดของคลังข้อมูลภาษาที่มีจำนวนเพิ่มมากขึ้นอย่างมหาศาลทำให้มีความจำเป็นต้องประยุกต์ใช้คอมพิวเตอร์และแนวคิดทางสถิติเพื่อช่วยสกัดข้อมูลต่างๆ ที่เหมาะสมต่อการวิเคราะห์ศัพท์

บทที่ 6 คลังข้อมูลกับการประมวลผลภาษา

1. บทนำ

ประโยชน์ที่สำคัญอีกประการหนึ่งของคลังข้อมูลภาษา คือการใช้คลังข้อมูลภาษาเป็นพื้นฐานในการฝึกสอน (train) ระบบการประมวลผลภาษา (Natural Language Processing) ให้เข้าใจลักษณะบางประการของภาษา ทั้งนี้ประโยชน์ที่ได้จากการใช้คลังข้อมูลภาษาในลักษณะนี้คือการได้ข้อมูลทางสถิติในแง่มุมต่างๆของการใช้ภาษา สถิติการปรากฏร่วมกันของคำต่างๆ จากนั้นจึงนำค่าสถิติที่หาได้จากคลังข้อมูลภาษามาประยุกต์ใช้ในงานทางด้าน NLP ต่อไป ที่ผ่านมา ระบบ NLP ที่ประยุกต์ใช้ข้อมูลทางสถิตินี้ประสบความสำเร็จมากกว่าระบบ NLP แบบดั้งเดิมที่เน้นในเรื่องการหาและประยุกต์ใช้กฎเกณฑ์ต่างๆทางภาษา ความสำเร็จทางด้านนี้เป็นสาเหตุสำคัญอันหนึ่งที่ทำให้ภาษาศาสตร์คลังข้อมูลเป็นที่สนใจอย่างมากในปัจจุบันนี้ และยังกล่าวได้ว่า แนวคิดของการประยุกต์ใช้สถิติเป็นกระแสหลักของสาขาภาษาศาสตร์คอมพิวเตอร์ในปัจจุบันอีกด้วย

2. ความสัมพันธ์ระหว่างภาษาศาสตร์และสถิติ

Abney (1996) กล่าวว่า นักภาษาศาสตร์โดยทั่วไปดูจะไม่รู้ถึงกระแสความเป็นไปที่เกิดขึ้นในสาขาด้านภาษาศาสตร์คอมพิวเตอร์ หรือถ้ารู้ ก็มักจะมองด้วยสายตาในเชิงลบ คิดว่าสถิติไม่น่าจะเกี่ยวข้องอะไรกับภาษาศาสตร์ โดยมองว่าการที่นักวิจัยทาง NLP ใช้สถิติจากข้อมูลภาษาจริงเพื่อฝึกสอนระบบ NLP

ของเขาให้รู้จักกับภาษานั้นๆ เป็นเพียงทางออกง่ายๆที่จะทำให้ระบบนั้นใช้งานได้ เป็นเพียงวิธีการแก้ปัญหาของพวกนักคอมพิวเตอร์ที่ไม่รู้จักภาษาศาสตร์ดีพอ ซึ่งอาจเป็นไปได้ว่าคนพวกนี้ไม่เคยศึกษาวิชาทางภาษาศาสตร์มาก่อน แต่หากมองอีกด้าน ก็อาจเป็นไปได้ว่า พวกนักคอมพิวเตอร์ได้พบว่างานทางภาษาศาสตร์นั้นไม่สามารถประยุกต์ใช้ได้โดยตรงต่อการพัฒนาระบบ NLP เขาจึงต้องหาวิธีอื่นๆ ที่จะทำให้งานด้านนี้เป็นไปได้มากขึ้น

นอกจาก Abney จะชี้ให้เห็นว่านักภาษาศาสตร์โดยทั่วไปมักไม่สนใจกับเรื่องของสถิติ และไม่ได้ตระหนักถึงงานของนักภาษาศาสตร์คอมพิวเตอร์ Abney ยังกล่าวเสริมอีกว่า โดยธรรมชาติแล้ว ความคิดของคนสองกลุ่มนี้จะต่างกัน ในมุมมองแบบสุดขั้วของนักภาษาศาสตร์คอมพิวเตอร์ นักภาษาศาสตร์คอมพิวเตอร์มองว่า ถ้าเราป้อนข้อมูลภาษาจำนวนมากเพียงพอให้กับระบบ NLP ให้ได้เรียนรู้ลักษณะของภาษาจากข้อมูลนั้น ระบบนั้นก็จะสามารถประมวลผลภาษาธรรมชาติได้โดยไม่ต้องอาศัยความรู้ทางภาษาศาสตร์ แต่จากมุมมองอีกสุดขั้วหนึ่งของนักภาษาศาสตร์ทั่วไป นักภาษาศาสตร์ก็จะแย้งว่าถึงแม้ระบบแบบที่ใช้สถิตินั้นจะทำงานได้ดีเพียงใด สิ่งที่ได้ก็ไม่ได้อธิบายอะไรเกี่ยวกับกลไกการเข้าใจภาษาของมนุษย์ซึ่งเป็นจุดมุ่งหมายหลักของภาษาศาสตร์ อย่างไรก็ตาม Abney มองว่า เรื่องทางสถิตินั้นไม่ใช่ว่าจะไม่เกี่ยวข้องกับทางภาษาศาสตร์เสียทีเดียว แต่มีส่วนเกี่ยวข้องที่จะช่วยให้เราเข้าใจถึงกลไกของภาษาได้ดีขึ้น โดย Abney ได้ชี้ให้พิจารณาเรื่องทางภาษาอย่างน้อย 3 เรื่อง คือ การรับภาษา (language acquisition) การเปลี่ยนแปลงของภาษา (language change) และการแปรของภาษา (language variation) ว่าเป็นเรื่องที่มีส่วนเกี่ยวข้องโดยตรงกับเรื่องสถิติ

1. การรับภาษา (language acquisition)

Abney กล่าวว่า ถ้าเด็กเรียนรู้ไวยากรณ์ (grammar) ในลักษณะแบบพีชคณิต (algebraic) คือแบบที่มีลักษณะเป็นกฎเกณฑ์แน่นอนตายตัวว่าอะไรใช่หรือไม่ใช่ ถูกหรือผิด เราคงจะเห็นการเปลี่ยนแปลงของการใช้ภาษาของเด็กอย่างเฉียบพลัน กฎบางอย่างจะเกิดขึ้นและกฎบางอย่างจะหายไปในวันนี้วันพรุ่งนี้ แต่ในความจริง การเปลี่ยนแปลงทางภาษาของเด็กเป็นการเปลี่ยนแปลงแบบค่อยเป็นค่อยไป เป็นการเปลี่ยนแปลงในเชิงความถี่สัมพัทธ์ทางโครงสร้าง ซึ่งน่าจะเป็นว่า เด็กจะมีกฎของหลายๆ โครงสร้างที่แข่งขันกันอยู่ภายใน สิ่งที่กำหนดว่าโครงสร้างไหนถูกต้องมากกว่าอาจจะมองในลักษณะที่เป็นค่าความน่าจะเป็น (probability) ซึ่งเปลี่ยนแปลงค่าไปได้เรื่อยๆในระหว่างการลองผิดลองถูก (try out) จนกระทั่งโครงสร้างใดที่มีค่าความน่าจะเป็นเป็นศูนย์ เด็กก็จะเลิกใช้โครงสร้างนั้นไป ลักษณะนี้เป็นการเพิ่มเอาสถิติในเรื่องความน่าจะเป็นเข้าไปในส่วนของไวยากรณ์ (grammar) ทำให้ไวยากรณ์มีลักษณะที่เป็นแบบสถิติ (stochastic) แทนที่จะเป็นแบบพีชคณิต (algebraic)

2. การเปลี่ยนแปลงของภาษา (language change)

ในเรื่องการเปลี่ยนแปลงของภาษานั้นก็เป็นเช่นเดียวกัน ถ้ากฎทางภาษามีลักษณะเป็นแบบพีชคณิตเราก็คงจะเห็นการเปลี่ยนแปลงทางภาษาอย่างทันทีทันใด แต่ความจริงไม่ได้เป็นเช่นนั้น การเปลี่ยนแปลงของภาษาเป็นเรื่องที่ใช้เวลาเป็นสิบเป็นร้อยปี ไม่ใช่ว่า อยู่ๆเราเดินไปที่ร้านเหล้าแห่งหนึ่ง แล้วสั่ง ale (ออกเสียง /el/) ซึ่งเป็นเบียร์ชนิดหนึ่งกลับได้ eel (ออกเสียง /il/) หรือปลาไหลมาแทนเพราะว่าเมื่อวานนี้ ระหว่างที่เราออกนอกเมืองไปได้เกิดการเลื่อนสระในภาษานี้จาก /e/ เป็น /i/ แต่ถ้าหากเรามองว่า ภาษาในกลุ่มสังคมหนึ่งๆเป็นภาพรวมเชิงสถิติของภาษาของแต่ละบุคคลในสังคมนั้นๆ เช่น ค่าความน่าจะเป็น

ของโครงสร้างหนึ่งในสังคมนั้นจะเท่ากับสัดส่วนของคนที่ใช้โครงสร้างนั้นต่อคนทั้งหมด ในมุมมองแบบนี้ เราก็สามารถมองการเปลี่ยนแปลงของภาษาในลักษณะที่เป็นการเปลี่ยนแปลงในลักษณะที่ค่อยๆ เลื่อนออกจากศูนย์กลางของคนในสังคมนั้นๆได้ คือจากที่เป็นที่นิยมใช้กันแพร่หลาย ก็ค่อยๆลดลงตามเวลาที่เปลี่ยนไป

3. การแปรของภาษา (language variation)

ในการแปรของภาษานั้นสามารถพิจารณาได้ในสองแบบ คือ การแปรในเรื่องภาษาถิ่น (dialectology) และการแปรในเรื่องแบบลักษณ์ภาษา (typology) ซึ่งทั้งสองแบบนี้ก็สามารถอธิบายผ่านแบบจำลองทางสถิติได้ กล่าวคือความแตกต่างของภาษาถิ่นก็เป็นเรื่องของความแตกต่างที่เกิดจากการใช้โครงสร้างแบบต่างๆในสัดส่วนที่ต่างกัน (mix frequency of usage of various constructions) ส่วนการแปรในเรื่องของแบบลักษณ์ภาษานั้นสามารถมองในเชิงสถิติได้ คือแทนที่จะมองการแปรระหว่างภาษาในลักษณะสมบูรณ์แต่ให้มองความเหมือนหรือความแตกต่างระหว่างภาษาต่างๆ ในเชิงสัดส่วนของสากลลักษณ์ (universal feature) ที่แตกต่างกัน

แต่คำถามที่ต้องถามต่อไปก็คือ แล้วสถิติเกี่ยวข้องกับไวยากรณ์ของบุคคลแต่ละคนที่อยู่ในสังคมที่พูดภาษาเดียวกัน (monolingual speaker in homogeous speech community) หรือไม่ เพราะนี่คือบริบทที่ Chomsky อ้างถึงในการสร้างแบบจำลองของสามัตถิยะภาษา (language competence) ถ้าพิจารณาอย่างผิวเผิน คำตอบก็คือ ไม่น่าจะเกี่ยวข้องกัน เรื่องทางสถิติไม่เกี่ยวข้องกับสิ่งที่นักภาษาศาสตร์ส่วนใหญ่กำลังศึกษาอยู่ ซึ่งเป็นเรื่องของการหากฎเกณฑ์ (principle) ที่อยู่ภายในของ language faculty ซึ่งเป็นแบบจำลองในอุดมคติ (idealization) ของคนที่อยู่ในสังคมที่พูดภาษาที่เหมือนกัน แต่ Abney

ได้ตั้งข้อสงสัยว่าสิ่งที่นักภาษาศาสตร์สนใจนี้แคบเกินไปหรือเปล่า ในสภาพการณ์ปัจจุบันนี้ แบบจำลองของไวยากรณ์ที่นักภาษาศาสตร์สร้างขึ้นสามารถอธิบายภาษาได้ในขอบเขตจำกัดเท่านั้น ข้อมูลอื่นๆที่ไม่สามารถอธิบายได้หรืออยู่นอกเหนือจากกฎเกณฑ์มักจะถูกมองว่าเป็นเรื่องของกฤตกรรมภาษา (performance) ไม่ใช่เรื่องของสามัตถิยะภาษา (competence) และออกตัวว่าอยู่นอกเหนือจากขอบเขตของการศึกษาทางภาษาศาสตร์ กล่าวคือ นักภาษาศาสตร์ไม่ได้ให้ความสำคัญกับข้อมูลจริงมากนัก แต่กลับสนใจอยู่กับข้อมูลสังเคราะห์ซึ่งไม่พบเห็นในชีวิตจริง และบางครั้งก็ยากที่ทุกคนจะตัดสินออกมาในลักษณะเดียวกันว่าเป็นประโยคที่ถูกไวยากรณ์หรือไม่ เช่น ในประโยคต่อไปนี้

?วันนี้ ผมเห็นคน$_i$ที่แดงบอกหน่อยว่าดำกำลังหาหนังสือที่ e$_i$ แต่ง

Whom$_i$ do you think that Bill said that John thought that Harry shaved e$_i$?

*Who$_i$ did you say was bothered by our talking to e$_i$?

และในความเป็นจริง มีข้อมูลมากมายที่ไวยากรณ์แบบที่ใช้กัน คือไวยากรณ์ในเชิงพีชคณิตไม่สามารถอธิบายได้ จริงๆแล้วคุณสมบัติแบบพีชคณิตอาจจะเป็นเพียงคุณสมบัติหนึ่งในหลายๆ คุณสมบัติของภาษา เป็นไปได้ไหมว่าเราอาจจะกำลังมองข้ามแง่มุมบางอย่างของภาษาไป Abney กล่าวสรุปในเรื่องนี้ว่านักภาษาศาสตร์ได้หลงลืมจุดมุ่งหมายเดิม และการเดินตามแนวคิดของ Chomsky เป็นการปิดหนทางการศึกษาภาษาอย่างแท้จริง (linguistics has lost sight of its original goal, and turned Chomsky's expedient into an end in itself)

การประยุกต์ใช้ไวยากรณ์แบบพีชคณิตในระบบการประมวลผลภาษาธรรมชาติทำให้เราเริ่มเห็นว่า ไวยากรณ์แบบพีชคณิตที่ใช้กันนั้นอาจไม่

สอดคล้องกับการรับรู้ของมนุษย์ และเป็นสาเหตุที่ทำให้การประมวลผลภาษาเป็นเรื่องยาก กล่าวคือในประโยคหนึ่งๆที่วิเคราะห์นั้นจะมีโครงสร้างที่เป็นไปได้มากมาย ก่อให้เกิดปัญหาในเรื่องของความกำกวมซึ่งระบบประมวลผลภาษาธรรมชาติ (NLP) จะต้องตัดสินว่าโครงสร้างใดประโยคแบบใดควรจะเป็นโครงสร้างที่ต้องการ หากลองพิจารณาตัวอย่างประโยคจาก Manning and Schutze (1999: 17) ประโยคอย่าง *Companies are training workers* ซึ่งคนเรารับรู้และเข้าใจ (perceive) แบบเดียวไม่สะดุดว่ามีความกำกวม คือ ในแบบที่ *are training* เป็นกลุ่มของกริยา แต่ระบบประมวลผลภาษาสามารถมองเห็นเพิ่มอีกสองแบบ คือ แบบที่ *are* เป็นกริยาหลักและส่วนที่เหลือเป็น gerund เหมือนอย่างในประโยค *Our problem is training workers* หรือแบบที่ *are* เป็นกริยาหลักส่วน *training* เป็นส่วนขยายของ *workers* เหมือนอย่างในประโยค *Those are training wheels* หรือในตัวอย่างประโยค *List the sales of the products produced in 1973 with the products produced in 1972* (Manning and Schutze 1999: 18) ซึ่งเป็นประโยคธรรมดาที่เราอาจใช้สั่งงานคอมพิวเตอร์ ในประโยคนี้ มีผู้รายงานว่า ระบบประมวลผลภาษาธรรมชาติของเขาสามารถวิเคราะห์แจงส่วน (parse) ได้ถึง 455 แบบ การพยายามที่จะแก้ปัญหานี้หรือการพยายามลดความกำกวมที่เกิดขึ้น โดยใช้วิธีเพิ่มกฎทางไวยากรณ์ก็จะส่งผลกระทบต่อประโยคอื่นๆ ทำให้ความสามารถในการวิเคราะห์หรือแจงส่วนประโยคอื่นๆลดลง ในขณะที่ถ้าพยายามเพิ่มกฎทางไวยากรณ์ให้ครอบคลุมประโยคแบบต่างๆมากขึ้นก็จะส่งผลให้ระบบประมวลผลภาษาธรรมชาติมองเห็นความกำกวมเพิ่มมากตามไปด้วย ปัญหาเหล่านี้ เราจะพบมากเวลาที่แจงส่วนประโยคต่างๆ ประโยคที่ดูเหมือนเป็นประโยคง่ายๆและไม่มีความกำกวมทางโครงสร้าง แต่ระบบประมวลผลภาษาธรรมชาติจะพบว่ามี

โครงสร้างที่เป็นไปได้มากกว่าหนึ่งโครงสร้าง คำถามคือ จะทำอย่างไรที่จะทำให้ระบบประมวลผลภาษาธรรมชาตินั้นสามารถเลือกโครงสร้างให้สอดคล้องกับแบบที่คนเรารับรู้ได้ วิธีการหนึ่งก็คือ การนำเรื่องของความน่าจะเป็นเข้ามาประยุกต์ใช้ โดยกำหนดให้โครงสร้างแต่ละโครงสร้างมีค่าความน่าจะเป็นที่ไม่เท่ากัน โครงสร้างที่มีค่าความน่าจะเป็นสูงเกินค่าระดับหนึ่ง จึงจะถือว่าเป็นโครงสร้างที่คนเราสามารถรับรู้ได้

ปัญหาเหล่านี้ชี้ให้เห็นว่าไวยากรณ์ในแบบพีชคณิตที่ศึกษากันอยู่นั้นไม่สอดคล้องกับการรับรู้ (perception) ของมนุษย์ คำถามหนึ่งที่เรามักถามตัวเองเสมอในเวลาที่ต้องตัดสินความถูกผิดของภาษาคือการถามว่าประโยคหรือวลีนี้ฟังดูเป็นธรรมชาติไหม (sound natural) ซึ่งคำตอบที่ได้จากคำถามในลักษณะนี้เป็นเรื่องที่เกี่ยวข้องกับระดับชั้น (degree) เราไม่สามารถอธิบายได้ด้วยกฎที่มีลักษณะตายตัวแยกแยะแบบเด็ดขาด (discrete) หรือที่เรียกว่ากฎแบบพีชคณิตได้ แต่อาจอธิบายได้ด้วยการใช้ไวยากรณ์ที่มีการถ่วงน้ำหนัก (weighted grammar) หรือไวยากรณ์เชิงสถิติ (stochastic) คือ ไวยากรณ์ที่ใช้ค่าความน่าจะเป็นประกอบด้วย ซึ่งแบบจำลองของไวยากรณ์แบบนี้น่าจะสอดคล้องกับวิจารณญาณของมนุษย์ (human judgment) มากกว่า กล่าวคือ ประโยคที่สร้างขึ้นจากไวยากรณ์นั้นนอกจากจะบอกได้ว่าถูกไวยากรณ์หรือไม่แล้ว ยังบอกถึงระดับชั้น (degree) ว่าฟังดูเป็นธรรมชาติเพียงใด และก็ควรจะต้องสอดคล้องกับการรับรู้ของมนุษย์

ตัวอย่างข้างล่าง (Abney 1996:13) เป็นเพียงตัวอย่างสมมติที่แสดงให้เห็นถึงการผนวกคุณสมบัติทางสถิติเข้าไปในไวยากรณ์แบบ context free grammar เพื่อทำให้เป็นไวยากรณ์ที่มีการถ่วงน้ำหนัก โดยมีการให้น้ำหนักหรือค่าความน่าจะเป็นของกฎแต่ละกฎไว้ข้างท้าย ซึ่งจะช่วยแก้ปัญหาในเรื่องความ

กำกวมทางโครงสร้างที่เกิดขึ้นได้

1. S -> NP V 0.7
2. S -> NP 0.3
3. NP -> N 0.8
4. NP -> N N 0.2
5. N -> John 0.6
6. N -> walks 0.4
7. V -> walks 1.0

จากไวยากรณ์ที่กำหนดให้นี้ *John walks* มีโครงสร้างที่เป็นไปได้อยู่สองแบบ คือ เป็น S หรือ NP แต่ด้วยน้ำหนักของแต่ละกฎที่แตกต่างกัน ทำให้โครงสร้าง S มีน้ำหนักเท่ากับ .7 x .8 x .6 x 1.0 = .336 ส่วนโครงสร้าง NP มีน้ำหนักเท่ากับ .3 x .2 x .6 x .4 = 0.0144 เนื่องจากน้ำหนักของโครงสร้าง S สูงกว่าโครงสร้าง NP มากจึงถือได้ว่าเป็นโครงสร้างที่รับรู้ (perceive) ได้ และเราอาจตั้งสมมติฐานต่อว่า ประโยคจะถูกมองว่ากำกวมถ้าอัตราส่วนระหว่างตัวเลขสูงสุดกับตัวเลขถัดไปไม่แตกต่างกันมาก เช่น ไม่เกิน 2:1 ความจริงแล้วแนวคิดเรื่องระดับชั้นของความถูกต้องทางไวยากรณ์ (degree of grammaticality) ก็มีปรากฏใช้กันอยู่ในทฤษฎีทางภาษาศาสตร์ เช่น การที่มองว่าการละเมิดของกฎ Empty Category Principle (violation of ECP) รุนแรงกว่าการละเมิดของกฎ Subjacency Principle (ทั้งสองเป็นกฎที่อยู่ในทฤษฎีไวยากรณ์ Government and Binding ของ Chomsky) แต่คำถามที่เกิดตามมาจากการผนวกคุณสมบัติของสถิติเข้ากับตัวไวยากรณ์ก็คือ ค่าตัวเลขความน่าจะเป็นเหล่านี้จะได้มาจากไหน คำตอบที่เป็นไปได้ คือ ตัวเลขพวกนี้เรียกว่าตัวเลขที่ได้จากการประมาณค่า (parameter estimation) ซึ่งน่าจะได้มาในระหว่างการเรียนรู้ภาษานั่นเอง

นอกจากนี้ Abney ยังเสริมให้เห็นว่าการใช้สถิตินั้นสามารถใช้อธิบายเรื่องของการปรากฏร่วมกันของคำ (collocation) และข้อจำกัดของการเลือกเกิดร่วมกัน (selectional restriction) ได้ หรืออย่างน้อยก็บอกถึงระดับชั้นของความเป็นธรรมชาติ (degree of naturalness) ได้ เช่น บอกได้ว่า *strong tea* และ *powerful car* จะฟังดูเป็นธรรมชาติมากกว่า *strong car* กับ *powerful tea* เพราะจากค่าทางสถิติมีการปรากฏร่วมกันของ *strong* กับ *tea* มากกว่า *strong* กับ *car* ซึ่งบางคนอาจจะแย้งว่าเป็นการยากที่จะกำหนดว่า ความเป็นธรรมชาติ (naturalness) คืออะไร แต่ Abney แย้งว่าในความเป็นจริงความเป็นธรรมชาติเป็นสิ่งที่เราสามารถรับรู้ได้ ความเป็นธรรมชาติเป็นคนละเรื่องกับการมีความหมายที่เข้าใจได้ (meaningfulness) ตัวอย่างเช่น เรารู้สึกว่า *differential structure* ฟังดูเป็นธรรมชาติมากกว่า *differential child* ถึงแม้ว่าเราจะไม่รู้ว่า *differential structure* หมายถึงอะไรก็ตาม นอกจากนี้ Abney ยังกล่าวว่าเรื่องทางสถิติยังสามารถนำมาใช้อธิบายปรากฏการณ์เรื่องอื่นๆในทางภาษาได้อีกด้วย ตัวอย่างเช่น

(a) การแก้ไขข้อผิดพลาดได้ (error tolerance) เช่น เวลาที่ผู้ฟังได้ยินประโยค *Thanks for all you help* ผู้ฟังมักเลือกที่จะตีความประโยคนี้ใหม่เป็น *thanks for all your help* มากกว่าที่จะคิดถึงประโยคนี้ในโครงสร้างที่มีหมายความว่า *thanks for all those who you help* ซึ่งก็เป็นโครงสร้างที่ถูกไวยากรณ์ เรื่องนี้สามารถอธิบายได้ว่า เพราะความสิ้นเปลือง (cost) ของการแก้ข้อผิดพลาด (error correction) เพื่อให้ได้โครงสร้างที่รู้สึกดีกว่า (preferred) (ซึ่งมีค่าสถิติการใช้มากกว่า) นั้นน้อยกว่าความสิ้นเปลืองของการพยายามสร้าง (derive) โครงสร้างที่ถูกไวยากรณ์แต่มีการค่าสถิติการใช้น้อยกว่า (less preferred)

(b) การเรียนรู้ในทันที (learning on the fly) เช่น เวลาที่ผู้ฟังได้ยินประโยคว่า *a hectare is a hundred ares* ผู้ฟังจะสรุปว่า *ares* เป็นคำนามได้โดยที่ไม่จำเป็นต้องรู้จักคำนี้มาก่อน การที่คนเราสามารถสร้างโครงสร้างประโยคที่คิดเอาไว้ได้ (pick intended structure) สามารถอธิบายได้ว่า คนเรามีกระบวนการที่จะเรียนรู้ โดยการกำหนดหมวดคำให้กับคำใหม่ เพิ่ม subcategorization frame ของคำกริยาที่เกี่ยวข้อง และคำนวณค่าความสิ้นเปลืองของการเรียนรู้นั้น ซึ่งการใช้ไวยากรณ์แบบถ่วงน้ำหนักได้ หรือไวยากรณ์ที่มีการใช้ค่าความน่าจะเป็นก็เป็นวิธีหนึ่งที่สามารถนำมาอธิบายปรากฏการณ์นี้ได้

นอกจากนี้ แนวคิดทางสถิติยังช่วยอธิบายในเรื่องของการเลือกโครงสร้างตามความนิยมใช้ (structural preference) ได้ เช่น ในประโยค *the emergency crews hate most is domestic violence* ในประโยคนี้นั้นโครงสร้างที่ถูกคือ *the emergency* เป็นนามวลีและ *crews hate most* เป็นส่วนขยายของนามวลี *(the emergency (that) crews hate most)* แต่เนื่องจากการเลือกตีความแบบที่ยาวที่สุดก่อนเป็นทางเลือกที่มักใช้มากกว่า ในการอ่านรอบแรก เราจึงอ่านนามวลีตัวแรกจบลงที่ *the emergency crews* แล้วมอง *hate* เป็นคำกริยา จนเมื่ออ่านต่อๆไปจึงพบว่าทางเลือกแบบนี้ไม่ถูกต้อง แล้วค่อยย้อนกลับไปเลือกอีกโครงสร้างหนึ่ง

3. การใช้สถิติในการแก้ปัญหาความกำกวม

สถิติพื้นฐานเรื่องความน่าจะเป็น สามารถนำมาประยุกต์ใช้แก้ปัญหาความกำกวมที่เกิดขึ้นในการประมวลผลภาษาธรรมชาติได้เป็นอย่างดี ในส่วนนี้เป็นการนำเสนอตัวอย่างหนึ่งของการใช้สถิติมาช่วยแก้ปัญหาเรื่องความกำกวมที่พบในการประมวลผลภาษาธรรมชาติ หลักการที่สำคัญในเรื่องนี้ คือ การนำ

แนวคิดของทฤษฎีความน่าจะเป็นมาใช้ ซึ่งแนวคิดทางด้านทฤษฎีความน่าจะเป็นนั้น มีแนวคิดโดยสรุปดังนี้

สำหรับเหตุการณ์ใดใด ค่าความน่าจะเป็นจะอยู่ระหว่าง 0 ถึง 1 เหตุการณ์ e1 จะไม่มีโอกาสเกิดขึ้น ถ้าค่าความน่าจะเป็นเป็นศูนย์ แต่จะเกิดขึ้นอย่างแน่นอน ถ้าค่าความน่าจะเป็นเป็นหนึ่ง ค่าความน่าจะเป็นของเหตุการณ์ใดใดคำนวณได้จากสถิติของการเกิดในเหตุการณ์ทั้งหมดที่ผ่านมา ตัวอย่างเช่น ในการแข่งม้า 100 เที่ยว ม้าชื่ออาภาพรชนะ 15 ครั้ง ถือได้ว่า ค่าความน่าจะเป็นที่ม้าอาภาพรจะชนะอีกคือ 0.15 และค่าความน่าจะเป็นที่ม้าอาภาพรจะแพ้เป็น 0.85 และถ้าในการแข่งขันเราพิจารณาเรื่องสภาพอากาศด้วย เช่น ในการแข่งม้า 100 ครั้งนี้ มีฝนตก 20 ครั้ง ซึ่งใน 20 ครั้งนี้ ม้าอาภาพรชนะ 10 ครั้ง ดังนั้นโอกาสที่ม้าอาภาพรจะชนะเมื่อฝนตกจะเป็น 0.5 (10 จาก 20 ครั้ง) ความน่าจะเป็นในแบบหลังนี้เรียกว่าความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ซึ่งเขียนแทนด้วย P(A | B) หมายถึงค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ A หลังจากที่เกิดเหตุการณ์ B แล้ว ซึ่งสามารถคำนวณได้โดยใช้สูตรต่อไปนี้

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

ในตัวอย่างข้างต้น P(อาภาพรชนะ | ฝนตก) = P(อาภาพรชนะและฝนตก) / P(ฝนตก) = (10/100) / (20/100) = 0.5 นอกจากนี้แล้ว ยังมีอีกสูตรหนึ่งที่เกี่ยวข้องกับความน่าจะเป็นแบบมีเงื่อนไข คือ กฎของ Bayes ซึ่งเขียนได้ดังนี้

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

จากตัวอย่างนี้ เราสามารถคำนวณค่า P(ฝนตก | อาภาพรชนะ) ได้ = P(อาภาพรชนะ | ฝนตก) * P(ฝนตก) / P(อาภาพรชนะ) = 0.5 * 0.2 / 0.15 = 0.67 หรือเมื่อคำนวณอีกแบบตามสูตรแรกจะได้ว่า P(ฝนตก | อาภาพรชนะ) = P(อาภาพรชนะ

และฝนตก) / P(อาภาพรชนะ) = 0.10 / 0.15 = 0.67 เท่ากัน จึงสามารถคำนวณโดยใช้สูตรใดก็ได้ตามแต่ข้อมูลที่มีอยู่

แนวคิดเรื่องความน่าจะเป็นนี้ถูกนำมาประยุกต์ใช้กับงานทางด้าน NLP ได้มากมาย แต่ในที่นี้ จะขอยกตัวอย่างเฉพาะการใช้ค่าความน่าจะเป็นในการแก้ความกำกวมเรื่องการกำกับหมวดคำ (part of speech tagging)[1] ว่าสามารถทำได้อย่างไร เพื่อให้เห็นแนวทางที่เป็นอยู่ในปัจจุบัน ในการกำกับหมวดคำให้กับคำทุกๆคำในประโยคนั้น จะพบว่า เมื่อพิจารณาเฉพาะคำๆเดียว คำบางคำสามารถเป็นได้มากกว่าหนึ่งหมวดคำ เช่น คำว่า flies เป็นได้ทั้ง N และ V ปัญหาที่เราต้องตัดสินคือ คำว่า *flies* นี้ควรเป็น N หรือ V ถ้าเขียนในรูปแบบของความน่าเป็น คือ เราต้องพิจารณาเปรียบเทียบระหว่าง PROB(C=N | W=flies) (อ่านว่า ค่าความน่าจะเป็นที่เมื่อกำหนดคำว่า flies ให้แล้วคำนั้นจะมีหมวดคำเป็น N) และ PROB(C=V | W=flies) แต่ในที่นี้จะเขียนย่อๆเป็น PROB(N | flies) และ PROB(V | flies) วิธีที่ง่ายที่สุดในการประยุกต์ใช้เรื่องความน่าจะเป็น คือ ดูว่าค่าความน่าจะเป็นของหมวดคำใดมากกว่ากัน ก็ให้เลือกหมวดคำนั้นเป็นคำตอบ ซึ่งจากกฎในเรื่องความน่าจะเป็น เราจะได้

PROB(N | flies) = PROB(flies & N) / PROB(flies)

PROBN(V | flies) = PROB(flies & V) / PROB(flies)

ในที่นี้ จะเห็นว่าค่าความน่าจะเป็นทั้งสองค่ามีตัวหารเหมือนกันคือ PROB(files) ดังนั้นคำตอบของปัญหานี้จึงขึ้นอยู่กับการตัดสินว่า PROB(flies & N) หรือ PROB(flies & V) ที่มากกว่ากัน ค่าตัวหารไม่มีผลใดใดต่อคำตอบ แต่

---

[1] ตัวอย่างและรายละเอียดในส่วนนี้เรียบเรียงจากบทที่ 7 ในหนังสือ Natural Language Understanding ของ Allen (1995)

เราจะหาค่าความน่าจะเป็นทั้งสองนี้ได้อย่างไร ในความเป็นจริงเราไม่สามารถกระทำได้ เพราะเราไม่สามารถเก็บรบวรวมประโยคทั้งหมดที่เคยมีการใช้กันมาในโลกนี้ได้ แต่ถ้าเรามีคลังข้อมูลภาษาขนาดใหญ่มากพอสมควร เราก็สามารถประมาณค่าความน่าจะเป็นอันนี้ได้จากคลังข้อมูลภาษานั้น เช่น ถ้ามีคลังข้อมูลภาษาขนาด 1,273,000 คำ และจากคลังข้อมูลพบว่ามีการใช้ *flies* 1,000 ครั้ง โดยใช้เป็น N 400 ครั้ง ใช้เป็น V 600 ครั้ง ดังนั้นเราจะได้

PROB(flies) ประมาณเท่ากับ 1,000/1,273,000 = 0.0008

PROB(flies & N) ประมาณเท่ากับ 400/1,273,000 = 0.0003

PROB(flies & V) ประมาณเท่ากับ 600/1,273,000 = 0.0005

ดังนั้น ถ้าให้เดา ระบบ NLP ก็ควรจะเดาว่า *flies* เป็น V เพราะมีโอกาสถูกร้อยละ 60 การประมาณแบบนี้เรียกว่าการประมาณค่าแบบมากสุดตามที่พบ (maximum likelihood estimator - MLE) คือคำนวณจากสัดส่วนจริงที่ปรากฏในคลังข้อมูลนั้น ความถูกต้องของค่าที่ประมาณได้นั้นขึ้นอยู่กับขนาดของข้อมูลที่ใช้ในการประมาณค่าความน่าจะเป็นเป็นสำคัญ ตัวอย่างเช่น ถ้าเราใช้การโยนเหรียญ 2 ครั้งมาช่วยในการประมาณค่าความน่าจะเป็นของการออกหัว ถ้าโยน 2 ครั้งได้หัว (H) หมด PROB(H) ที่ประมาณได้จะเท่ากับ 1 ถ้าโยน 2 ครั้งได้ก้อย (T) หมด PROB(H) ที่ประมาณได้จะเท่ากับ 0 ถ้าโยน 2 ครั้งได้ T หนึ่งครั้ง H หนึ่งครั้ง PROB(H) ที่ประมาณได้จะเท่ากับ 0.5 ถ้าสมมติว่าค่าประมาณของ PROB(H) ที่เรายอมรับว่าใช้ได้ในที่นี้เป็นค่าที่อยู่ระหว่าง .25 ถึง .75 การโยนเหรียญ 2 ครั้งจะให้โอกาสที่ค่าประมาณการจะใช้ไม่ได้ถึง 50% แต่เมื่อโยน 3 ครั้ง โอกาสที่ค่าประมาณของ PROB(H) จะไม่เป็นที่ยอมรับจะลดลงเป็น 25% ดูข้อมูลด้านล่างประกอบ (Allen 1995:193-194)

Results	Estimate of Prob(H)	AcceptableEstimate?
HH	1.0	NO
HT	0.5	YES
TH	0.5	YES
TT	0	NO

ตาราง 6.1 การประมาณค่า Prob(H) จากการโยนเหรียญ 2 ครั้ง

Results	Estimate of Prob(H)	Acceptable Estimate?
HHH	1.0	NO
HHT	0.66	YES
HTH	0.66	YES
HTT	0.33	YES
THH	0.66	YES
THT	0.33	YES
TTH	0.33	YES
TTT	0	NO

ตาราง 6.2 การประมาณค่า Prob(H) จากการโยนเหรียญ 3 ครั้ง

ตัวอย่างนี้แสดงให้เห็นว่าค่าประมาณการจะใช้ได้ดีก็ต่อเมื่อข้อมูลที่ใช้ในการประมาณนั้นมีขนาดใหญ่มากๆ ในกรณีของภาษา เราสามารถประมาณค่าความน่าจะเป็นได้ดีมากขึ้นถ้าหากว่าคลังข้อมูลภาษาที่ใช้มีขนาดใหญ่มากๆ แต่ถึงกระนั้น ข้อมูลภาษาก็ยังมีปัญหาเรื่องข้อมูลที่ปรากฏน้อย (sparse data) ทั้งนี้เพราะคำแต่ละคำมีความถี่ในการใช้ไม่เท่ากัน จะมีคำเป็นจำนวนมากที่มีความถี่ในการใช้ต่ำมาก และจะมีคำจำนวนน้อยที่มีความถี่ในการใช้สูงมาก ยิ่งกว่านั้นคำบางคำก็อาจจะไม่พบในคลังข้อมูลภาษาที่ใช้เลยก็ได้ ค่าความน่าจะเป็นของคำนั้นจึงเป็น 0 เมื่อค่าความน่าจะเป็นของการพบคำนั้นเป็น 0 ค่าความน่าจะเป็นของประโยคที่มีคำนั้นอยู่ก็จะเป็น 0 ไปด้วย จึงต้องมีวิธีประมาณค่าความน่าจะเป็นที่ดีกว่าแบบ MLE การประมาณค่าแบบคาดเดา (expected likelihood estimator - ELE) เป็นวิธีการหนึ่ง ซึ่งแก้ปัญหานี้ได้โดยการบวกตัวเลขน้อยๆ เช่น 0.5 ให้กับความถี่ในการเกิดของทุกคำ ดังนั้นจึงไม่มีคำใดที่มีค่าความน่าจะเป็นเป็น 0 แต่วิธีนี้ก็ไม่ใช่การแก้ปัญหาในเรื่องนี้ดี มีวิธีการแบบอื่นๆที่

ดีกว่า แต่จะไม่ขอกล่าวในที่นี้ ผู้สนใจสามารถศึกษาเพิ่มเติมได้จาก Jurafsky and Martin (2000) ในหัวข้อเรื่องของการปรับค่า (smoothing)

ในทางปฏิบัติ เมื่อใช้คลังข้อมูลภาษาหนึ่งในการประมาณค่าความน่าจะเป็นของคำต่างๆ เราควรมีคลังข้อมูลภาษาอีกคลังหนึ่งเพื่อใช้ทดสอบการทำงานของโปรแกรมที่ใช้ค่าความน่าจะเป็นที่ประมาณได้จากคลังข้อมูลแรก ไม่ควรทดสอบโปรแกรมบนคลังข้อมูลภาษาเดิมเพราะจะเป็นเหมือนการจำคำตอบมาตอบคำถาม การทดสอบบนคลังข้อมูลที่สองจะชี้ให้เห็นว่า ค่าความน่าจะเป็นที่ประมาณได้จากคลังข้อมูลแรกให้ผลเป็นที่น่าพอใจเพียงใด

การให้โปรแกรมตัดสินหมวดคำของคำแต่ละคำในประโยคโดยอัตโนมัติ วิธีที่ง่ายสุด คือวิธีแบบที่ได้กล่าวมาแล้วในตอนต้น คือการเลือกหมวดคำที่ให้ค่าความน่าจะเป็นของคำๆนั้นสูงสุด จากการศึกษาพบว่าวิธีนี้สามารถแก้ปัญหาความกำกวมของหมวดคำได้ถูกต้องประมาณ 90% ดังนั้น ค่าความถูกต้อง 90% นี้จึงควรเป็นค่าต่ำที่สุดที่ควรจะเป็น โปรแกรมที่ใช้วิธีการทางสถิติแบบใดใดก็ตามไม่ควรจะให้ค่าความถูกต้องน้อยไปกว่าค่านี้

วิธีการทางสถิติแบบหนึ่งที่ใช้กันทั่วไป คือการใช้บริบทข้างเคียงหรือคำที่ปรากฏข้างเคียงมาช่วยตัดสิน เช่น *flies* ที่ตามหลังคำ *the* ก็ควรจะเป็น N มากกว่า V หากเรามองปัญหาการกำกับหมวดคำ (POS tagging) นี้ให้เป็นปัญหาแบบที่เกี่ยวข้องกับความน่าจะเป็นไปได้ เราสามารถมองปัญหานี้ได้ว่า ถ้ากำหนดให้ประโยคหนึ่งๆ คือลำดับของคำ $W_1, W_2,\ldots W_t$ สิ่งที่เราต้องการคือหาลำดับของ $C_1,C_2,\ldots C_t$ ที่ให้ค่า

(1) $PROB(C_1,\ldots C_t \mid W_1,\ldots W_t)$ มากที่สุด

แต่ว่า การหาค่าประมาณการนี้โดยตรงจากคลังข้อมูลภาษานั้นจะต้องใช้ข้อมูลจำนวนมหาศาล เพราะโอกาสที่จะพบประโยคเดียวกันในคลังข้อมูลนั้นมี

น้อยมาก และโอกาสที่จะพบลำดับของหมวดคำแบบที่เป็น $C_1,C_2,....C_t$ กับประโยคนั้นก็ยิ่งน้อยลงไปอีก แต่หากเราแปลงปัญหานี้โดยใช้กฎของ Bayes ปัญหาในข้อ (1) จะถูกแปลงออกมาเป็นปัญหาของ

(2) $PROB(C_1,...C_t) * PROB(W_1,....W_t \mid C_1,....C_t)) / PROB(W_1,....W_t)$

เนื่องจากในที่นี้ เราต้องการหา $C_1,....C_t$ ที่ให้ค่าความน่าจะเป็นที่มากที่สุด เราจึงสามารถตัดตัวหาร $PROB(W_1,....W_t)$ ทิ้งได้ เพราะไม่มีผลอะไรต่อการเปรียบเทียบ $C_1,....C_t$ แบบต่างๆ ดังนั้น เราจึงสามารถมองปัญหานี้ออกมาเป็นปัญหาของ

(3) $PROB(C_1,...C_t) * PROB(W_1,....W_t \mid C_1,....C_t)$

แต่ปัญหานี้ก็ยังแก้ได้ยากอยู่ดี เพราะว่าจะต้องใช้ข้อมูลมากมายเพื่อที่จะใช้ในการคำนวณค่าความน่าจะเป็นอันนี้ จึงต้องหาวิธีประมาณค่าความน่าจะเป็นอันนี้ใหม่เพื่อให้สามารถหาค่าได้จริงๆ ซึ่งเราสามารถทำได้โดยตั้งข้อสันนิษฐานบางอย่างขึ้นมา ถึงแม้ว่าจะยังไม่มีการพิสูจน์ว่าข้อสันนิษฐานนี้เป็นจริงหรือไม่ แต่ในทางปฏิบัติที่ผ่านมา ก็ใช้ได้ผลเป็นอย่างดี ข้อสันนิษฐานที่ว่านี้คือ ค่าความน่าจะเป็นที่จะพบลำดับของหมวดคำแบบหนึ่งๆ ประมาณว่าเท่ากับผลคูณรวมของค่าความน่าจะเป็นที่จะพบหมวดคำในแต่ละตำแหน่งเมื่อพิจารณาเฉพาะหมวดคำก่อนหน้าจำนวนจำกัด เช่น 1 หรือ 2 ตำแหน่งข้างหน้า คือแทนที่จะหาค่าความน่าจะเป็นที่จะพบหมวดคำของคำทั้งหมดในประโยคตามลำดับที่เกิด ก็ให้คำนวณจากผลคูณรวมของค่าความน่าจะเป็นของหมวดคำทีละคำ โดยที่ค่าความน่าจะเป็นของหมวดคำแต่ละคำนั้นขึ้นอยู่กับหมวดคำที่ปรากฏข้างหน้าในขอบเขตจำกัด เช่น 1 หรือ 2 ตัวก่อนหน้าด้วย ข้อสันนิษฐานนี้เรียกว่า Markov assumption หรือที่เรียกกันว่าเป็นการใช้แบบจำลองแบบ bigram หรือ trigram เป็นต้น สมมติว่าในที่นี้เลือกใช้ bigram คือพิจารณาเพียงหนึ่งตำแหน่งข้างหน้า

ค่าความน่าจะเป็นที่ประมาณใหม่ได้จะเป็น

$$PROB(C_1,..., C_T) \cong \prod_{i=1,T} PROB(C_i \mid C_{i-1})$$

กล่าวคือ PROB(C_1,...C_t) ประมาณว่าเท่ากับ ผลคูณรวมของ i=1,t PROB(C_i | C_{i-1}) เช่น PROB(ART N V N) ประมาณว่าเท่ากับ PROB(ART | ϕ) * PROB(N | ART) * PROB(V | N) * PROB(N | V)

ทำนองเดียวกัน ตัวคูณที่สองในสมการ (3) ก็สามารถประมาณค่าใหม่โดยตั้งข้อสันนิษฐานว่า ค่าความน่าจะเป็นในการปรากฏของรูปคำ W สำหรับหมวดคำ C ใดใด หรือ PROB(W | C) ไม่ได้ขึ้นอยู่กับรูปคำและหมวดคำที่เกิดข้างหน้าหรือข้างหลัง ดังนั้น

PROB(W_1,...W_t | C_1,...C_t) จึงประมาณว่าเท่ากับ ผลคูณรวมของ PROB(W_i | C_i) โดยที่ i=1 ถึง t

เมื่อประมาณได้ดังนี้ ปัญหาใน (3) จึงมองได้ใหม่ว่าเป็นปัญหาของ

$$\prod_{i=1,T} PROB(C_i \mid C_{i-1}) * PROB(W_i \mid C_i)$$

หรือคือผลคูณรวมของ PROB(C_i | C_{i-1}) * PROB (W_i |C_i) โดยที่ i=1,t ซึ่งค่าเหล่านี้เราสามารถหาได้จากการนับความถี่ของคำและหมวดคำในคลังข้อมูลภาษา เช่น สมมติว่ามีคลังข้อมูลภาษาอังกฤษ ซึ่งมีหมวดคำที่เป็นไปได้อยู่ 4 หมวดคำ คือ N V ART และ P หากต้องการคำนวณค่าความน่าจะเป็นของการพบ V ก่อนหน้า N ก็คำนวณได้ดังนี้

ค่า PROB(C_i=V | C_{i-1}=N) ≈ Count(N ที่ตำแหน่ง i-1 และ V ที่ตำแหน่ง i) / Count(N ที่ตำแหน่ง i-1) ส่วน PROB(W_i | C_i) คำนวณจากการเกิดของคำต่างๆ ในหมวดคำนั้นๆ เช่น PROB(the | ART) = Count(จำนวนครั้งที่ the เป็น ART)/Count(จำนวนครั้งที่พบ ART) (ใช้กฎ PROB(A|B) = PROB(A&B) /PROB(B)) (ดูตารางที่ 6.3, 6.4 ประกอบ)

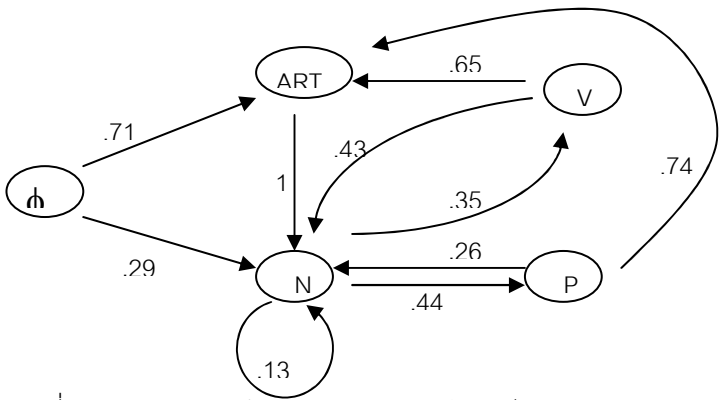
ในที่นี้ สมมติว่าต้องการหาหมวดคำของคำทุกคำในประโยค *flies like a flower* ถ้าใช้วิธีแบบง่ายๆแต่เสียเวลามากคือให้ทดลองสร้างลำดับ (sequence) ของหมวดคำ $C_1,...C_t$ ทุกแบบที่เป็นไปได้เพื่อหา $C_1,...C_t$ แบบที่ให้ค่าความน่าจะเป็นที่สูงที่สุด ซึ่งวิธีนี้จะต้องคำนวณค่าความน่าจะเป็นทั้งหมด N^T แบบ โดยที่ N คือจำนวนของหมวดคำ ส่วน T คือจำนวนของคำในประโยค แต่เราสามารถทำได้ดีกว่าโดยใช้แบบจำลอง probabilistic finite state machine มาช่วย ซึ่งสามารถแสดงเป็นรูปภาพได้ดังในรูปที่ 6.1 (ซึ่งค่าความน่าจะเป็นต่างๆ นั้นได้มาจากค่าในตารางที่ 6.3) ดังนั้น เราจึงสามารถคำนวณค่าความน่าจะเป็นเฉพาะของลำดับที่เป็นไปได้ตามแผงผังซึ่งจะมีจำนวนน้อยกว่า N^T อย่างแน่นอน ตัวอย่างเช่น ค่าความน่าจะเป็นของชุดที่เป็น N V ART N จะเท่ากับ .29 * .43 *.65 * 1 = .081

Category	Count at i	Pair	Count at i,i+1	Bigram	Estimate
0	300	0,ART	213	PROB(ART | 0)	.71
0	300	0,N	87	PROB(N | 0)	.29
ART	558	ART,N	558	PROB(N | ART)	1
N	833	N,V	358	PROB(V | N)	.43
N	833	N,N	108	PROB(N | N)	.13
N	833	N,P	366	PROB(P | N)	.44
V	300	V,N	75	PROB(N | V)	.35
V	300	V,ART	194	PROB(ART | V)	.65
P	307	P,ART	226	PROB(ART | P)	.74
P	307	P,N	81	PROB(N | P)	.26

ตารางที่ 6.3: แสดงความน่าจะเป็นของ bigram จากคลังข้อมูล (Allen 1995:198)

	N	V	ART	P	TOTAL
Flies	21	23	0	0	44
Fruit	49	5	1	0	55
Like	10	30	0	21	61
A	1	0	201	0	202
The	1	0	300	2	303
Flower	53	15	0	0	68
Flowers	42	16	0	0	58
Birds	64	1	0	0	65
Others	592	210	56	284	1142
TOTAL	833	300	558	307	1998

ตารางที่ 6.4 : แสดงความถี่ของคำในหมวดคำต่าง ๆ (Allen 1995: 199)

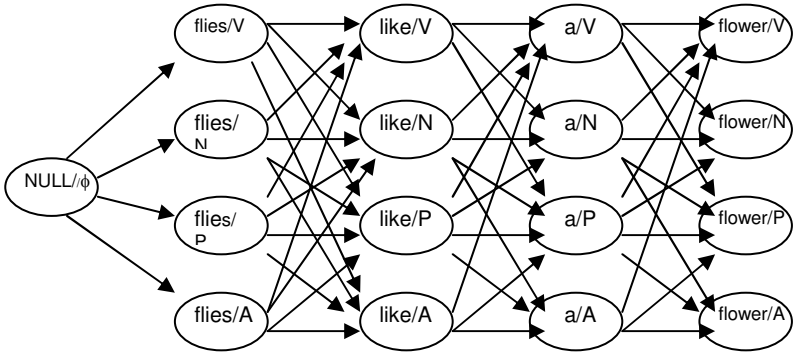


รูปที่ 6.1 ผัง Markov พร้อมค่าความน่าจะเป็นของไบแกรม (Allen 1995: 200)

เมื่อคำนวณค่า PROB(C_i | C_{i-1}) ได้แล้ว ขั้นต่อไปคือการคำนวณหาค่า PROB(W_i | C_i) ตัวอย่างเช่น ค่า PROB(W_i | C_i) ของประโยค *flies like a flower* เมื่อถูกกำหนดหมวดคำเป็น N V ART N ตามลำดับ จะคำนวณได้จาก PROB(flies | N) * PROB(like | V) * PROB(a | ART) * PROB(flower | N) ซึ่งสามารถคำนวณค่าความน่าจะเป็นที่ได้จากตารางที่ 6.4 ได้ผล = 0.25 * .1 * .36 * .063 = 5.4 * 10^{-5} เมื่อคูณด้วยค่า .081 ที่ได้จาก PROB(N V ART N) ก็จะได้

ค่าประมาณของความน่าจะเป็น ที่ N V ART N จะเป็นหมวดคำของ *flies like a flower* ตามลำดับ

ในกรณีที่วิเคราะห์ประโยคที่มีคำ 4 คำและมีหมวดคำที่เป็นไปได้ทั้งหมด 4 หมวดคำ ถ้าทำแบบง่ายๆ ก็ต้องคำนวณหาลำดับของหมวดคำที่เป็นไปได้ทั้งหมดแล้วเลือกทางที่ให้ค่าความน่าจะเป็นที่สูงที่สุด ซึ่งในที่นี้ต้องคำนวณทั้งสิ้น 4^4 = 256 (ดูรูปที่ 6.2) แต่เราอาจไม่จำเป็นต้องหาค่าของการแทนทุกลำดับชุดที่เป็นไปได้ เพราะเราอาจทำโดยการพิจารณาไปทีละคำ ในระหว่างนั้น จากลำดับ (sequence) ทั้งหมดที่เป็นไปได้ ให้จัดเป็นกลุ่มตามหมวดคำของคำสุดท้าย เช่น กลุ่มที่คำสุดท้ายเป็น N กลุ่มที่คำสุดท้ายเป็น V เป็นต้น จากนั้นให้เลือกลำดับที่ดีที่สุดในแต่ละกลุ่มนั้นมาเพื่อนำไปใช้สำหรับการพิจารณาคำถัดไป ทำเช่นนี้ไปเรื่อยๆจนครบทุกคำในประโยค วิธีการนี้เป็นการใช้ข้อสันนิษฐานของ Markov คือหมวดคำที่จะคิดต่อไปนั้นขึ้นอยู่กับหมวดคำก่อนหน้าเท่านั้น ดังนั้น ในบรรดาลำดับหมวดคำทั้งหมดที่คำสุดท้ายเป็นหมวดคำเดียวกันเราสามารถเลือกมาเพียงหนึ่งแบบ คือแบบที่มีค่าความน่าจะเป็นสูงที่สุด กล่าวคือ ในการคำนวณจริงจะเริ่มจากหาลำดับที่ดีที่สุด 4 แบบสำหรับ flies like ก่อน คือ หาลำดับที่ดีที่สุดที่คำสุดท้ายเป็น N ลำดับที่ดีที่สุดที่คำสุดท้ายเป็น ART ลำดับที่ดีที่สุดที่คำสุดท้ายเป็น V และลำดับที่ดีที่สุดที่คำสุดท้ายเป็น P จากนั้นหาลำดับที่ดีที่สุดสำหรับ flies like a ในหมวดคำสุดท้ายต่างๆ 4 แบบ ทำเช่นนี้ไปเรื่อยๆจนจบประโยค จึงไม่จำเป็นต้องคำนวณถึง 256 ครั้ง คือจะทำการคำนวณเพียง k*T*N2 ครั้งแทนที่จะเป็น N^T (โดยที่ k เป็นค่าคงที่ตัวหนึ่ง) ด้วยวิธีการเช่นนี้ (ซึ่งเรียกว่า Viterbi algorithm) เราจึงสามารถใช้ค่าความน่าจะเป็นที่ได้จากคลังข้อมูลภาษามาแก้ปัญหาความกำกวมที่เกิดขึ้นในปัญหาการกำหนดหมวดคำในประโยคได้อย่างมีประสิทธิภาพมากขึ้น



รูปที่ 6.2 แสดงความเป็นไปได้ของลำดับหมวดคำ 256 แบบ (Allen 1995:201)

3. บทสรุป

ในบทนี้ ได้แสดงให้เห็นถึงความสำคัญของคลังข้อมูลภาษาที่มีต่องานงานประมวลผลภาษาธรรมชาติ ซึ่งได้อาศัยคลังข้อมูลภาษาเป็นข้อมูลในการคำนวณค่าทางสถิติต่างๆ เกี่ยวข้องกับภาษา ค่าสถิติต่างๆนี้ จะถูกนำมาใช้ในงานที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ ซึ่งวิธีการแบบที่ใช้สสถิตินี้ถือเป็นกระแสหลักของงานการประมวลผลภาษาในปัจจุบัน เพราะจะเห็นว่าเป็นวิธีการที่มีความยืดหยุ่นสูงสามารถใช้แบบจำลองเดียวกันกับภาษาต่างๆ ได้ เพียงแค่เปลี่ยนคลังข้อมูลภาษาที่ใช้ในการฝึกสอนระบบเท่านั้น แนวคิดในเรื่องสถิติหรือความน่าจะเป็นนั้น นักภาษาศาสตร์บางกลุ่มก็ไม่ได้คิดว่าแป็นเรื่องแปลกปลอมของภาษาศาสตร์ แต่เป็นหลักการพื้นฐานหนึ่งของภาษาที่มนุษย์เราได้ใช้ความคุ้นเคยหรือประสบการณ์จากการใช้ภาษาจริงในการตัดสินเรื่องต่างๆ เกี่ยวกับภาษา โดยที่กฎไวยากรณ์ในภาษาไม่ได้มีลักษณะเป็นกฎแยกขาดชัดเจนแบบที่รู้จักกัน

บทที่ 7 คลังข้อมูลภาษาผู้เรียน

1. บทนำ

คลังข้อมูลภาษาผู้เรียน (learner corpus) เป็นคลังข้อมูลภาษาประเภทหนึ่งที่เก็บข้อมูลภาษาที่ได้จากผู้เรียนภาษาต่างประเทศ อาจเป็นภาษาเขียนหรือเป็นภาษาพูดก็ได้ เนื่องจากเป็นข้อมูลจากผู้ที่ไม่ได้เป็นเจ้าของภาษาจึงมีลักษณะแตกต่างจากคลังข้อมูลภาษาปกติ(ที่มักเก็บข้อมูลภาษาจากเจ้าของภาษา)ในหลายลักษณะ เป็นข้อมูลที่จำแนกตามช่วงอายุของผู้พูดได้ไม่ชัดเจนเท่าคลังข้อมูลภาษาปกติเพราะความสามารถในการใช้ภาษานั้นขึ้นกับประสบการณ์ภาษามากกว่าอายุของผู้พูด เป็นข้อมูลที่สะท้อนความไม่สมบูรณ์มากกว่าจะเป็นข้อมูลใช้อ้างอิง เป็นข้อมูลที่แปรไปตามกลุ่มผู้เรียนชาติต่างๆ ได้ด้วย และคลังข้อมูลภาษาผู้เรียนก็เป็นข้อมูลจริงที่ทำให้เราเข้าใจสภาพการใช้ภาษาที่ผู้เรียนภาษานั้นเป็นตลอดจนปัญหาที่เป็นอยู่ว่ามีความเบี่ยงเบนจากภาษามาตรฐานหรือภาษาที่เจ้าของภาษาใช้กันอย่างไร อะไรที่ใช้แล้วดูไม่เป็นธรรมชาติเหมือนเจ้าของภาษา ในขณะเดียวกัน ก็ทำให้เห็นภาษาในระหว่าง (interlanguage) ที่เกิดขึ้นในระหว่างพัฒนาการการเรียนรู้ภาษาต่างประเทศและภาษาที่สอง คลังข้อมูลภาษาผู้เรียนจึงเป็นทรัพยากรที่เป็นประโยชน์อย่างยิ่งต่อการเรียนการสอนภาษาต่างประเทศและภาษาที่สองในด้านต่างๆ

คำภาษาอังกฤษที่ใช้เรียกคลังข้อมูลภาษาผู้เรียน คือ learner corpus ซึ่งมีความหมายหมายเชิงกว้างหมายถึงข้อมูลภาษาอะไรก็ได้ที่ผู้เรียนมีภาษาแม่เป็นอีกภาษาหนึ่ง แต่เมื่อเติมรายละเอียดว่าเป็นการเรียนภาษาอะไรและผู้เรียน

เป็นคนชาติใดหรือภาษาแม่คือภาษาอะไรแล้ว ก็จะพบว่ามีการใช้คำที่หลากหลาย ในที่นี้จะยกตัวอย่างเฉพาะข้อมูลจากผู้เรียนภาษาอังกฤษ รูปแบบคำต่างๆ ที่สำรวจพบว่ามีการใช้แตกต่างกัน ดังต่อไปนี้ (โดยที่ ___ แทนภาษาแม่ของผู้เรียนหรือที่ที่เก็บข้อมูลผู้เรียนหรือประเภทย่อย)

1) ___ corpus of learner English เช่น international corpus of learner English, Greek corpus of learner English, HKUST(Hong Kong University of Science and Technology) corpus of learner English
2) corpus of ___ learner English เช่น Corpus of Japanese Learner English, corpus of Chinese learner English
3) corpus of ___ learner of English เช่น Corpus of Chinese Learner of English
4) ___ learner English corpus เช่น Polish Learner English Corpus, Chinese Learner English Corpus (CLEC)
5) ___ English learner corpus เช่น Thai English learner corpus, Chinese-English learner corpus, business English learner corpus, adult English learner corpus,
6) English learner corpus of ____ เช่น English learner corpus of Turkish students' Writings,
7) ___ learner corpus of English เช่น Taiwanese Learner Corpus of English, SILS Learner Corpus of English
8) ___ learner of English Corpus เช่น Santiago University Learner of English Corpus
9) Corpus of English by ___ learners เช่น Corpus of English by

Japanese Learners

ในที่นี้ จะใช้คำ “คลังข้อมูลภาษาผู้เรียน” สำหรับ learner corpus แทนที่จะใช้ “คลังข้อมูลผู้เรียนภาษา” เพราะจะได้เน้นที่ตัวภาษาของผู้เรียนมากกว่าที่ตัวผู้เรียนภาษา เดิมผู้เขียนตั้งใจจะใช้คำ “คลังข้อมูลผู้เรียนภาษา” เพื่อจะได้ล้อไปกับ native (speaker) corpus ด้วยคำ “คลังข้อมูลเจ้าของภาษา” ได้ แต่เมื่อพิจารณาแล้ว การใช้คำส่วนใหญ่เป็นการพูดถึงตัว learner corpus เองมากกว่าจะพูดเทียบกับ native corpus การใช้คำ “คลังข้อมูลภาษาผู้เรียน” จึงน่าจะได้ความชัดกว่า อีกทั้งสอดคล้องกับที่ใช้คำ “คลังข้อมูลภาษา” (corpus) “คลังข้อมูลภาษาอังกฤษ” (English corpus) แต่ก็อาจมีปัญหาเมื่อพูดถึง English learner corpus เพราะถ้าใช้ “คลังข้อมูลภาษาของผู้เรียนภาษาอังกฤษ” ก็จะดูเยิ่นเย้อไม่กะทัดรัดเท่า “คลังข้อมูลผู้เรียนภาษาอังกฤษ” ดังนั้น ในบางบริบท ผู้เขียนจะใช้คำ “คลังข้อมูลผู้เรียนภาษา” ในความหมายเดียวกับ “คลังข้อมูลภาษาผู้เรียน” ส่วนคลังข้อมูลภาษาอังกฤษของผู้เรียนชาวไทยจะใช้คำอังกฤษว่า Thai Learner English corpus เพราะดูจะเป็นรูปแบบที่แพร่หลายและกระชับมากกว่า และคำไทยนั้นจะใช้ “คลังข้อมูลภาษาอังกฤษของผู้เรียนชาวไทย” เป็นหลัก แต่อาจใช้คำ “คลังข้อมูลผู้เรียนภาษาอังกฤษชาวไทย” เมื่อต้องการเทียบกับ คลังข้อมูลผู้เรียนภาษาอังกฤษชาติอื่นๆ

ความสนใจในคลังข้อมูลภาษาผู้เรียนในปัจจุบันก่อเป็นรูปธรรมขึ้น โดยมีการจัดประชุมวิชาการนานาชาติในปีค.ศ.1999 ที่ฮ่องกง โดย Sylviane Granger และ Joseph Hung ซึ่งเป็นกลุ่มคนที่ร่วมในโครงการ ICLE (International Corpus of Learner English) และก็มีการประชุมวิชาการตามมาอีกอย่างต่อเนื่องจนถึงปัจจุบัน Diaz-Negrillo and Fernandez-Dominguez (2006) ให้ภาพรวมของงานวิจัยด้านคลังข้อมูลภาษาผู้เรียนว่าเป็นงานที่

ผสมผสานวิธีการวิจัยจากงานทางด้านภาษาศาสตรคลังข้อมูลและงานทางด้านการเรียนการสอนภาษาต่างประเทศหรือภาษาที่สอง โดยมีการวิเคราะห์เชิงปริมาณแบบภาษาศาสตร์คลังข้อมูล มีการกำกับข้อมูลภาษา กำกับข้อผิดชัดเจน ในขณะเดียวกันก็อาศัยวิธีการวิเคราะห์เปรียบต่าง (contrastive analysis) และวิธีการวิเคราะห์ข้อผิด (error analysis) แบบเดียวกับงานด้านการสอนภาษาต่างประเทศหรือภาษาที่สองเพื่อหาปัญหาและความต่างจากภาษาของเจ้าของภาษา

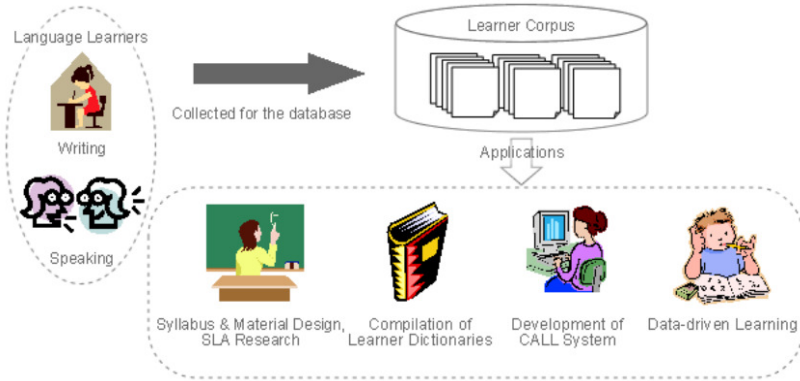
ความจริง การศึกษาวิจัยเรื่องการเรียนภาษาต่างประเทศที่ผ่านมาก็ได้อาศัยข้อมูลการใช้ภาษาของนักเรียนมาศึกษาข้อผิดต่างๆ แต่การใช้ข้อมูลที่ผ่านมาเป็นการใช้ข้อมูลที่เก็บมาแบบจำกัด ไม่สามารถให้คำตอบในภาพรวมได้ เช่น เมื่ออาจารย์ผู้สอนพบว่านักเรียนมีการใช้คำบางคำผิด ก็อาจจะมีคำถามตามมาว่าคำที่ใช้ผิดนี้ ในที่อื่นๆ หรืองานเขียนลักษณะอื่น นักเรียนใช้ผิดอย่างสม่ำเสมอหรือมีที่ใช้ถูกผสมด้วย ข้อผิดที่พบเป็นเรื่องเฉพาะตัวหรือเป็นปัญหาร่วมของนักเรียนคนอื่นๆ ด้วย และเมื่อนักเรียนมีความรู้ภาษาอังกฤษในระดับสูงขึ้น ข้อผิดนี้ยังคงพบอยู่หรือไม่ และข้อผิดนี้พบในงานเขียนของเจ้าของภาษาด้วยหรือไม่ เจ้าของภาษามีการใช้คำนี้อย่างไร แตกต่างจากผู้เรียนอย่างไร การที่จะตอบคำถามเหล่านี้ได้ต้องอาศัยข้อมูลมากกว่างานเขียนที่อาจารย์ผู้สอนเก็บจากนักเรียนของตนเองเท่านั้น แต่ต้องอาศัยข้อมูลภาษาจากผู้เรียนจำนวนมากและหลากหลายพอจึงจะตอบคำถามเหล่านี้ได้ชัดเจนขึ้น คลังข้อมูลภาษาผู้เรียนจึงมีบทบาทที่สำคัญยิ่งในการศึกษาวิจัยเรื่องการเรียนการสอนภาษาต่างประเทศ Granger (2004) กล่าวถึงคลังข้อมูลภาษาผู้เรียนที่สร้างกันในปัจจุบันว่ามีประเด็นที่ควรคำนึงดังต่อไปนี้

ขนาด คลังข้อมูลภาษาผู้เรียนจะมีขนาดที่ใหญ่มากกว่าการเก็บข้อมูลการใช้ภาษาของนักเรียนภายในชั้นเรียน โดยคลังข้อมูลภาษาผู้เรียนในปัจจุบันมีจำนวนคำเป็นหลักล้านคำ ทำให้มีความน่าเชื่อถือของการวิเคราะห์และตีความผลข้อมูลมากขึ้น ซึ่งฐานข้อมูล CHILDES (MacWhinney 2000) ที่ใช้กันแพร่หลายในการศึกษาการรับภาษาก็ถูกสร้างขึ้นด้วยเหตุผลนี้ แต่เรื่องขนาดเพียงอย่างเดียวก็ไม่เพียงพอสำหรับการศึกษาภาษาหากไม่มีการออกแบบการเก็บข้อมูลที่ดีพอ

การออกแบบ ต้องคำนึงตัวแปรต่างๆ ที่เกี่ยวข้องและมีผลต่อลักษณะข้อมูลภาษาผู้เรียนที่รวบรวมได้ ตัวอย่างเช่น ในคลังข้อมูล ICLE (International Corpus of Learner English) (Granger 2003) มีการกำหนดตัวแปรทั่วไปที่เกี่ยวกับผู้พูดเช่น อายุ เพศ ภาษาแม่ ตัวแปรที่เกี่ยวกับตัวบท เช่น ประเภทตัวบท หัวข้อเรื่อง ความยาว และตัวแปรเฉพาะสำหรับ L2 เช่น ประสบการณ์ภาษา ภาษาอื่นๆที่รู้ ตัวแปรเกี่ยวกับงาน เช่น เวลาที่ให้ สถานการณ์(เป็นการสอบการบ้าน) มีการใช้หนังสืออ้างอิงช่วย เป็นต้น

2. การใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียน

คลังข้อมูลภาษาผู้เรียนเป็นประโยชน์โดยตรงต่อการศึกษาวิจัยทางด้านภาษาศาสตร์และด้านการเรียนการสอนภาษาต่างประเทศ หากพิจารณาจากลักษณะการใช้ประโยชน์ในด้านต่างๆ ไม่ว่าจะเป็นการศึกษาวิจัยการเรียนการสอนภาษาต่างประเทศ เพื่อให้เข้าใจปัญหาและธรรมชาติของการเรียนภาษาต่างประเทศแล้วนำไปออกแบบการเรียนการสอนภาษาที่เหมาะสม ตลอดจนการพัฒนาทรัพยากรต่างๆ ให้เหมาะกับผู้เรียนชาวต่างประเทศ หรือนำไปพัฒนาระบบคอมพิวเตอร์ช่วยสอน หรือเป็นแหล่งข้อมูลให้นักเรียนได้เข้าถึงข้อมูลและเห็นปัญหาการใช้ภาษาด้วยตนเอง



Learner corpus and its applications

รูปจาก http://www2.nict.go.jp/x/x161/en/member/izumi_emi/project.html

2.1 ประโยชน์ในการศึกษาข้อผิด (error analysis)

การศึกษาข้อผิดเป็นการศึกษาที่มีมาตั้งแต่ในยุคแรกของการศึกษาวิจัยการเรียนการสอนภาษาต่างประเทศ เพราะเชื่อว่าจะทำให้เห็นปัญหาต่างๆ ในการเรียนภาษา มีการจำแนกประเภทข้อผิดเป็นแบบต่างๆ และมีการศึกษาข้อผิดเป็นจำนวนมากช่วงทศวรรษ 1970 แต่ก็มีข้อติติงเรื่องของการใช้ข้อมูลที่เป็นการเก็บจากข้อมูลจำนวนน้อยและการเก็บที่ไม่เป็นระบบ จึงไม่สามารถสะท้อนสถานภาพภาษาในระหว่าง (interlanguage) ของผู้เรียนได้ครอบคลุมและน่าเชื่อถือเพียงพอ ยากต่อการที่จะตรวจสอบใหม่หรือทำซ้ำ การจำแนกข้อผิดเองก็มีคำถามเรื่องความชัดเจนในการวิเคราะห์เพราะบางกรณีอาจตัดสินได้ยากว่าเป็นข้อผิดระดับคำ ระดับไวยากรณ์ หรือระดับการสะกดคำ การตัดสินก็มีลักษณะที่เป็นอัตวิสัยไม่มีมาตรฐานที่เป็นเกณฑ์ร่วมกัน และการสนใจมองเฉพาะสิ่งที่เป็นข้อผิด ก็ทำให้มองข้ามสิ่งที่ไม่ถูกจัดว่าเป็นข้อผิด การใช้ภาษาที่ถูกแล้วไม่มีการพูดถึง ไม่เห็นรูปแบบการใช้ภาษาที่มากเกินหรือน้อยเกินของผู้เรียน

(overuse และ underuse) เพราะผู้เรียนสามารถเลี่ยงการใช้คำหรือวลีที่ตนไม่มั่นใจเพื่อไม่ให้ปรากฏข้อผิดได้ (Dagneaux et.al. 1998) และเมื่อคลังข้อมูลภาษาเริ่มเป็นที่ยอมรับใช้กันแพร่หลายในทศวรรษที่ 1990 ความสนใจในภาษาผู้เรียนก็ทำให้มีการรวบรวมข้อมูลเป็นคลังข้อมูลภาษาผู้เรียนอย่างเป็นระบบ และการกำกับข้อผิดก็เป็นกระบวนการหนึ่งในการวิเคราะห์และใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียน ทำให้เห็นถึงข้อผิดได้ว่าอะไรเป็นข้อผิดที่พบมากในกลุ่มผู้เรียนกลุ่มต่างๆ

ตัวอย่างของงานที่ศึกษาข้อผิดโดยอาศัยคลังข้อมูลผู้เรียนภาษาอังกฤษ เช่น ข้อผิดเรื่องคำปรากฏร่วม (collocation) (Chi et al. 1994; Lorenz 1997; Granger 1998; Chen 1998) ข้อผิดเรื่องการใช้คำเชื่อม (connectors) (Milton and Tsang 1993; Granger and Tyson 1996; Satoh and Fang 1998) ข้อผิดเรื่องการใช้คำยกเว้นรูปอดีตกาล (irregular past tense) (Tono and Aoki 1998) ข้อผิดเรื่องการใช้คำนำหน้านาม (article) (Mason and Uzar 2000)

2.2 ประโยชน์ในการศึกษาเปรียบต่าง

การศึกษาเปรียบต่าง (contrastive analysis) แต่เดิมเป็นการศึกษาลักษณะความแตกต่างของภาษาที่หนึ่งและภาษาที่สอง เพื่อที่จะคาดการณ์ว่าปัญหาของการเรียนภาษาที่สองหรือภาษาต่างประเทศนั้นจะมีอะไรบ้าง เดิมจึงเป็นการศึกษาโดยไม่จำเป็นต้องอาศัยคลังข้อมูลภาษาแต่อาศัยความรู้ทางไวยากรณ์ของทั้งสองภาษา แต่ในปัจจุบัน การศึกษาเปรียบต่างเป็นไปในรูปแบบใหม่ที่อาศัยข้อมูลการใช้ภาษาจริงที่เกิดขึ้น และไม่ได้จำกัดที่การเปรียบต่างระหว่างภาษาต่างประเทศที่เรียนกับภาษาแม่ของผู้เรียน แต่มองเปรียบต่างภาษาเดียวกันได้ เช่น ภาษาอังกฤษของเจ้าของภาษากับภาษาอังกฤษของผู้เรียนชาวต่างชาติ Granger เรียกว่าเป็นการวิเคราะห์เปรียบต่างภาษาในระหว่าง

(Contrastive Interlanguage Analysis) เมื่อใช้คลังข้อมูลผู้เรียนภาษากับคลังข้อมูลเจ้าของภาษา ก็จะสามารถศึกษาพัฒนาการการเรียนภาษาที่หนึ่งเทียบกับภาษาที่สองในแง่มุมต่างๆ ตั้งแต่การใช้คำ รูปประโยค คำปรากฏร่วม วัจนลีลา ว่ามีอะไรที่แตกต่างจากภาษาอังกฤษต้นแบบอย่างไรบ้าง โดยมีพื้นความคิดว่าความแตกต่างเหล่านี้คือปัญหาของการเรียนภาษาอังกฤษ แต่ก็มีคนตั้งคำถามเรื่องว่าภาษาอังกฤษต้นแบบคืออะไร เป็นแบบอังกฤษ หรือแบบอเมริกัน หรือแบบอื่นๆ เพราะหลายๆ ประเทศในปัจจุบันก็ใช้ภาษาอังกฤษเป็นภาษาประจำชาติ บางคน (เช่น Tan 2005) จึงเสนอว่าลักษณะภาษาที่พบในคลังข้อมูลผู้เรียนไม่ได้เป็นเรื่องของข้อผิดเสมอ แต่รูปแบบการใช้ที่ไม่เหมือนกับเจ้าของภาษามีสาเหตุจากวัฒนธรรมของผู้พูดเองก็ได้

ตัวอย่างงานที่ศึกษาเปรียบต่างภาษาอังกฤษของผู้เรียนกับของเจ้าของภาษา ได้แก่ Biber and Reppen (1998) ศึกษาเปรียบต่าง complement clause ใน Longman Grammar Corpus กับ Longman Learners' Corpus แม้ว่าผลจะพบว่ามีการใช้คล้ายๆกัน แต่ก็ยังต้องอาศัยงานที่มีข้อมูลผู้เรียนมากกว่านี้เพื่อยืนยันผลต่อไป

นอกจากจะเปรียบกับภาษาที่หนึ่งแล้ว ก็มีผู้ที่สนใจเปรียบเทียบภาษาอังกฤษของผู้เรียนที่มีภาษาแม่แตกต่างกัน เพื่อศึกษาว่าภาษาแม่ที่แตกต่างกันนั้นจะมีผลที่เหมือนหรือแตกต่างกันต่อการพัฒนาความรู้ภาษาอังกฤษของผู้เรียนหรือไม่ ซึ่งก็เป็นอีกความหมายหนึ่งของการวิเคราะห์เปรียบต่างภาษาในระหว่าง (Contrastive Interlanguage Analysis) ตามที่ Granger ใช้ ผลที่ได้จากการศึกษาเปรียบต่างแบบนี้ทำให้เราเห็นถึงข้อผิดที่เป็นสากล (universal errors) ไม่ว่าผู้เรียนจะมีภาษาแม่อะไรก็ตาม และทำให้เห็นถึงข้อผิดที่เป็นเรื่องเฉพาะของแต่ละภาษา (L1-specific errors) คลังข้อมูลภาษาผู้เรียนจากชาติ

ต่างๆ จึงเป็นประโยชน์โดยตรงต่อการเข้าใจกระบวนการการรับภาษาต่างประเทศ (foreign language acquisition) อย่างไรก็ตาม ในการเปรียบต่างข้อมูลผู้เรียนชาติต่างๆ ระดับการศึกษาเดียวกัน เช่น ระดับมัธยม ไม่ได้เป็นหลักประกันผู้เรียนแต่ละชาติจะมีความสามารถภาษาอังกฤษระดับที่สามารถเทียบกันได้ อาจจะต้องดูจากเกณฑ์อื่น เช่น ช่วงคะแนนสอบมาตรฐานอย่าง TOEFL แต่ก็ต้องพิจารณาด้วยว่าถ้าใช้การทดสอบแบบที่เป็นมาตรฐานนี้ ก็อาจทำให้ได้กลุ่มตัวอย่างที่น้อยลง เพราะมีค่าใช้จ่ายสำหรับการทดสอบนี้แก่กลุ่มตัวอย่างด้วย

2.3 ประโยชน์ในการศึกษาวิเคราะห์การใช้ภาษา

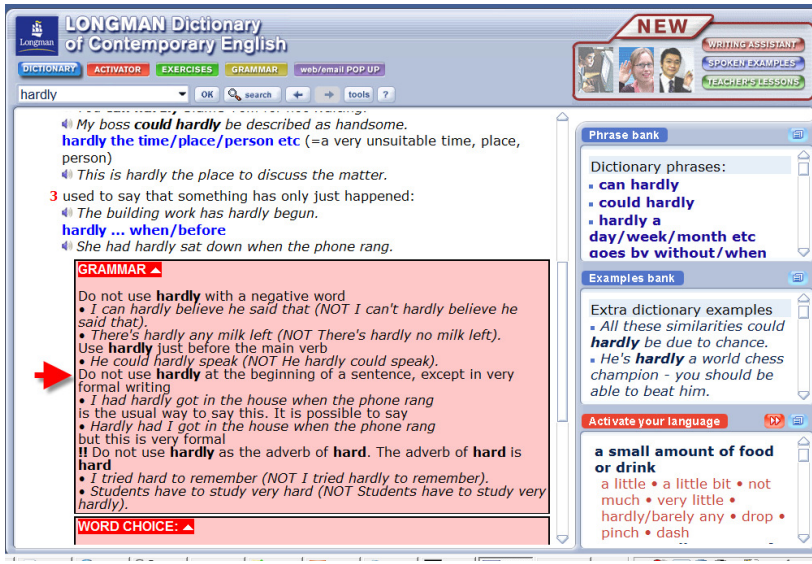
เป็นศึกษาวิเคราะห์การใช้ภาษาที่เกิดขึ้น (performance analysis) ของภาษาในระหว่าง (interlanguage) ว่ามีการเปลี่ยนแปลงหรือแตกต่างระหว่างกลุ่มต่างๆ อย่างไร หากจะศึกษาพัฒนาการของภาษาในระหว่างควรจะศึกษาจากคลังข้อมูลภาษาผู้เรียนในช่วงเวลาต่าง ๆ เช่น งานของ Leńko-Szymańska (2000a, b) ที่ศึกษาพัฒนาการจำนวนวงศัพท์ในผู้เรียนโดยศึกษาจากคลังข้อมูลภาษาผู้เรียน

2.4 ประโยชน์ด้านการจัดการเรียนการสอน

เป็นประโยชน์ในลักษณะที่มีต่อผู้เรียนภาษาอังกฤษ ในมิติแรกเป็นการนำผลที่ได้จากการศึกษาเปรียบต่าง การศึกษาข้อผิด และการศึกษาวิเคราะห์การใช้ภาษา มาเป็นพื้นฐานในการออกแบบการเรียนการสอนที่จะเหมาะสมกับผู้เรียนในแต่ละกลุ่มแต่ละช่วงอายุ หรือสร้างทรัพยากรเพื่อการเรียนการสอนไม่ว่าจะเป็นพจนานุกรม ตำรา หรือโปรแกรมคอมพิวเตอร์ช่วยสอน การจัดทำทรัพยากรเพื่อการเรียนการสอนภาษาอังกฤษเริ่มต้นโดยใช้คลังข้อมูลเจ้าของภาษา ไล่ตั้งแต่การจัดทำพจนานุกรม Collins Cobuild (1987) ภายใต้การนำของ

John Sinclair ตลอดจนถึงการทำหนังสือไวยากรณ์ A Comprehensive Grammar of the English Language (Quirk et al. 1985) และ Collins Cobuild English Grammar (Sinclair 1990) ซึ่งได้อาศัยตัวอย่างที่ได้จากคลังข้อมูลภาษา หนังสือไวยากรณ์อย่าง The Longman Grammar of Spoken and Written English (Biber et al. 1999) ซึ่งเขียนขึ้นโดยอาศัยความรู้จากการวิเคราะห์คลังข้อมูลภาษา Longman Spoken and Written English Corpus ขนาด 40 ล้านคำ ในทรัพยากรการเรียนการสอนยุคหลังจึงมีส่วนของการใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียน ตัวอย่างเช่น ในพจนานุกรม Longman Language Activator (1993), Longman Dictionary of Contemporary English (1995, 2003) และ Cambridge Advanced Learner's Dictionary พจนานุกรมเหล่านี้มีการให้ข้อมูลเพิ่มเติมในเรื่องที่ผู้เรียนมักใช้ภาษาอังกฤษผิดบ่อย (Granger 2004) เช่น คำว่า hardly เป็นคำที่พบว่าผู้เรียนมักมีปัญหาใช้ผิดจึงมีการให้ข้อมูลเพิ่มเติมในพจนานุกรมดังแสดงในรูปข้างล่าง

ในอีกมิติเป็นการให้ผู้เรียนได้ใช้คลังข้อมูลภาษาเพื่อการเรียนรู้ภาษาซึ่งโดยทั่วไปก็มักเป็นการใช้คลังข้อมูลของเจ้าของภาษา ที่ผู้เรียนจะสามารถสำรวจทำความเข้าใจการใช้ภาษาด้วยตนเอง แต่ในอีกลักษณะก็สามารถใช้คลังข้อมูลผู้เรียนได้ เพื่อให้เห็นรูปแบบข้อผิดที่พบบ่อยในคลังข้อมูลจะได้มีความตระหนักในปัญหาเหล่านั้น



รูป 7.1 แสดงคำอธิบายเพิ่มเติมคำที่ผู้เรียนมักใช้ผิดปรากฏแทรกในพจนานุกรม

ตัวอย่างงานที่ใช้ประโยชน์จากคลังข้อมูลผู้เรียนเพื่อช่วยสอนผู้เรียน เช่น งานของ Milton (1998) ที่ศึกษาปัญหาผู้เรียนชาวจีนที่พบจากคลังข้อมูลภาษาผู้เรียน แล้วพัฒนาโปรแกรมคอมพิวเตอร์ช่วยสอน AutoWord เพื่อให้คำแนะนำแก่ผู้เรียนในการเขียนความเรียง โดยภายในโปรแกรมประกอบด้วยส่วนที่รู้จำข้อผิด (error recognition) งานของ Mark (1998a, b) ได้พัฒนาคลังข้อมูลภาษาผู้เรียนจากงานเขียนของนักศึกษาในชั้นเรียนของเขา และนำมาใช้ประกอบกิจกรรมการเรียนการสอน

การศึกษาข้อมูลภาษาผู้เรียนโดยทั่วไปเป็นการศึกษาด้วยการใช้ศัพท์ไวยากรณ์ ปริจเฉท (lexis, grammar, and discourse) ในการเรียนภาษาที่สองซึ่งแสดงผลในรูปของการใช้เกิน ใช้ขาด และใช้ผิด (over use, under use, and misuse) ไม่เป็นธรรมชาติ ซึ่งเป็นการมองเมื่อเทียบกับการใช้ภาษาของเจ้าของ

ภาษา อย่างไรก็ดี Tan (2005) เสนอให้พิจารณาว่าภาษาผู้เรียนว่าเป็นภาษาในแบบของผู้เรียนเองซึ่งเป็นผลจากการที่ภาษาอังกฤษกลายเป็นภาษากลาง (lingual franca) ไปแล้วและได้รับอิทธิพลของวัฒนธรรมประเทศนั้นๆเอง ที่ทำให้แตกต่างจากภาษาอังกฤษ การศึกษาภาษาผู้เรียนจึงอาจไม่จำเป็นต้องเน้นเรื่องการเบี่ยงเบนจากมาตรฐานภาษาอังกฤษก็ได้

3. การพัฒนาคลังข้อมูลภาษาผู้เรียน

การออกแบบคลังข้อมูลเป็นเรื่องสำคัญและจะต้องให้ข้อมูลที่ชัดเจนว่าเก็บข้อมูลใดมาบ้าง เพราะจะทำให้ผู้ใช้เข้าใจลักษณะข้อมูลที่รวบรวมมา ซึ่งจะเป็นประโยชน์ในการศึกษาและวิเคราะห์เปรียบเทียบผลที่ได้ ก่อนเก็บข้อมูลผู้จัดทำจึงต้องมีการวางแผนออกแบบคลังข้อมูลที่ต้องการว่าควรมีลักษณะอย่างไร Tono (2003) แจกแจงว่าเราสามารถพิจารณาจากปัจจัยต่างๆ ดังนี้

ปัจจัยเรื่องของภาษาได้แก่ mode เป็นภาษาพูดหรือภาษาเขียน genre ประเภทของตัวบท เช่น เป็นความเรียง จดหมาย บันทึกประจำวัน เรื่องสั้น เป็นต้น style รูปแบบการเขียน เช่น เรื่องเล่า การอภิปราย topic หัวข้อเรื่อง เช่น เรื่องทั่วไป เรื่องนันทนาการ

ปัจจัยเรื่องชิ้นงาน (task-related) ได้แก่ การรวบรวมข้อมูลว่าเป็นการเก็บหลากหลาย (cross-selectional) หรือเก็บต่อเนื่อง (longtitudinal) วิธีการเก็บข้อมูลว่ามีการเตรียมการ (prepared) หรือเก็บตามสะดวก (spontaneous) การใช้ข้อมูลอ้างอิง (มีการใช้พจนานุกรม ใช้ต้นฉบับ) เวลาที่ใช้ (จำกัด ไม่จำกัด) ที่ที่ทำงาน (ทำในชั้นเรียน ทำเป็นการบ้าน)

ปัจจัยผู้เรียน ได้แก่ internal-cognitive (age, cognitive-style) internal-affective (motivation, attitude) L1 background L2 background (ESL หรือ EFL, และระดับชั้นศึกษา) L2 proficiency (standard test score)

ส่วนการกำกับข้อมูลขึ้นกับความต้องการว่าต้องการกำกับข้อมูลใดบ้าง ที่มีการกำกับในคลังข้อมูลภาษาผู้เรียน ได้แก่ header tagging เพื่อให้ข้อมูลต่างๆเกี่ยวกับตัวบทและสถานการณ์ error tagging เพื่อกำกับข้อผิดที่พบและประเภทของข้อผิด pos tagging เพื่อกำกับหมวดคำของแต่ละคำในข้อมูล sentence tagging เพื่อกำกับขอบเขตแต่ละประโยค lemmatization เพื่อบอกถึงรากคำแต่และคำ parsing เพื่อกำกับข้อมูลโครงสร้างประโยคว่าประกอบด้วยส่วนต่างๆอย่างไรบ้าง discourse tagging เพื่อกำกับข้อมูลปริจเฉท เช่น การทักทาย การขอโทษ แสดงความสุภาพ เป็นต้น, prosody tagging เพื่อกำกับข้อมูลสัทสัมพันธ์ anaphoric tagging เพื่อกำกับการอ้างอิงภายในตัวบท เช่น สรรพนามที่ใช้หมายถึงใคร

การกำกับข้อผิดนั้นก็มีความหลากหลายแล้วแต่ความต้องการและแนวทฤษฎีที่ใช้ แต่ก็มีบางโครงการที่พยายามสร้างมาตรฐานในกลุ่ม เช่น ICLE ซึ่งรวมข้อมูลภาษาผู้เรียนจากหลายๆ ประเทศ การพัฒนาโปรแกรมกำกับข้อผิดโดยอัตโนมัติยังเป็นไปได้ยากอยู่ไม่ว่าจะใช้แนวทางแบบกฎหรือแบบสถิติ ทั้งนี้เพราะยังไม่มีข้อมูลเรื่องข้อผิดที่เพียงพอว่ามีอะไรบ้างมากน้อยเท่าไร มีความต่างกันในกลุ่มผู้เรียนต่างกลุ่มอย่างไร การกำกับข้อผิดจึงเป็นไปในลักษณะทำด้วยคนมากกว่า แต่ก็เป็นงานที่หนักและใช้เวลามาก ควบคุมให้สม่ำเสมอยาก ในปัจจุบันจึงมีการศึกษาเรื่องการใช้โปรแกรมคอมพิวเตอร์ช่วยกำกับข้อผิด (Computer-aided Error Analysis หรือ CEA) เช่น Dagneaux (1998), Granger (2003), Diaz-Negrillo (2006) ซึ่งเป็นโปรแกรมเพื่อช่วยให้มีการกำกับข้อผิดได้ง่ายขึ้น เป็นระบบเพราะใช้แท็กต่างๆ ที่ออกแบบไว้แล้วว่าจะระบุข้อผิดอะไรได้บ้าง นอกจากนี้ ด้วยเหตุที่ข้อผิดที่พบในงานผู้เรียนสามารถแก้ไขได้ในหลายวิธี เช่น

The policeman was in this corner whistle but it was too late สามารถแก้ไขใหม่ได้สามแบบ

1 The policeman who was in this corner whistled but it was too late

2 The policeman was in this corner and whistled but it was too late

3 The policeman__in this corner whistled but it was too late

การกำกับข้อผิดว่าเป็นข้อผิดอะไรจึงเป็นงานที่เป็นอัตวิสัย จึงมีพยายามที่จะพัฒนาระบบที่ใช้ในการกำกับข้อผิดสำหรับงานในโครงการต่างๆ ซึ่งจำเป็นต้องออกแบบให้ง่ายต่อการใช้งานและช่วยให้การกำกับข้อผิดเป็นไปอย่างคงเส้นคงวา (consistent) เช่น งานของ Tono et.al. (2001) งานของ Dagneaux, Denness and Granger (1998) งานของ Izumi (2003)

โปรแกรมกำกับหมวดคำ ที่มีอยู่เป็นโปรแกรมสำหรับกำกับหมวดคำงานเขียนโดยเจ้าของภาษา การนำโปรแกรมเล่านี้มาใช้สำหรับงานเขียนนักเรียนต่างชาติจะมีปัญหาเรื่องความถูกต้องของการกำกับที่ต้องพิจารณาด้วย เพราะโปรแกรมกำกับหมวดคำภาษาอังกฤษที่ใช้กันพัฒนามาจากพื้นการใช้ภาษาอังกฤษของเจ้าของภาษา ซึ่งในการกำกับหมวดคำโปรแกรมที่เคยกำกับได้ถูกต้องมากกว่า 80% เมื่อมาใช้กับข้อมูลภาษาผู้เรียนที่มีลักษณะการเขียนที่ต่างมีข้อผิดทางไวยากรณ์ ความถูกต้องจึงน่าจะลดต่ำลงอีกมาก มีตัวอย่างงานที่ศึกษาผลจากการกำกับข้อมูลหมวดคำแล้วพบว่าข้อผิดของผู้เรียนทำให้ความถูกต้องการกำกับหมวดคำลดลงไป เช่น Van Rooy and Schäfer (2003) การตรวจสอบแก้ไขด้วยคนภายหลังก็เป็นงานที่ค่อนข้างใช้เวลาและสิ้นเปลือง

4. โครงการพัฒนาคลังข้อมูลภาษาผู้เรียนที่ต่างๆ

ปัจจุบันมีคลังข้อมูลภาษาผู้เรียนเกิดขึ้นเป็นจำนวนมาก ในที่นี้จะจำกัดเฉพาะคลังข้อมูลภาษาผู้เรียนภาษาอังกฤษ ซึ่งก็มีทั้งแบบที่พัฒนาโดยเอกชน

และที่พัฒนาโดยสถานศึกษา Granger (2004) สรุปโครงการต่างๆ ไว้ ส่วนที่เป็นคลังข้อมูลภาษาผู้เรียนเพื่อการพาณิชย์ ได้แก่ Longman Learners' Corpus และ Cambridge Learner Corpus โดยมีขนาด 10 ล้านคำและ 16 ล้านคำ ตามลำดับ ส่วนคลังข้อมูลที่พัฒนาโดยสถานศึกษามีขนาดที่แตกต่างกัน อย่าง ICLE ที่ออกมาในปีค.ศ. 2002 มีขนาด 2.5 ล้านคำ แต่จำนวนมากนั้นไม่ได้เผยแพร่แก่สาธารณะแต่มีไว้ใช้ภายในหน่วยงาน นอกจากนี้ คลังข้อมูลภาษาผู้เรียนที่จัดทำโดยสถานศึกษามักเป็นข้อมูลจากผู้พูดภาษาเดียว ในขณะที่คลังข้อมูลของ Longman และ Cambridge นั้นมีข้อมูลจากผู้พูดหลากหลายกว่า เพราะต้องการครอบคลุมผู้เรียนภาษาอังกฤษจากชาติต่างๆ จะมียกเว้นก็ ICLE ที่รวมกลุ่มนักวิจัยชาติต่างๆ เพื่อทำคลังข้อมูลภาษาผู้เรียนในชาติของตนและมีการแลกเปลี่ยนข้อมูลกัน ทำให้มีข้อมูลภาษาผู้เรียนจากหลากหลายชาติที่สามารถนำมาศึกษาเปรียบเทียบกันได้

โครงการ ICLE (International Corpus of English)[1] (Granger 2002) เป็นโครงการนานาชาติที่สร้างคลังข้อมูลผู้เรียนภาษาอังกฤษจากประเทศต่างๆ โดยกำหนดโครงสร้างและประเภทของตัวบทแบบเดียวกันเป็น argumentative essays ความยาว 500-1000 คำ มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบภาษาอังกฤษของผู้เรียนชาติต่างๆ ข้อมูลมาจากผู้เรียนภาษาอังกฤษระดับสูง (advanced learners) ซึ่งเป็นนักศึกษาสาขาวิชาภาษาอังกฤษชั้นปีที่ 3 หรือ 4 โครงการนี้เริ่มต้นในปีค.ศ.1990 โดย Granger University of Louvain-la Neuve ประเทศ Belgium ในปีค.ศ.2002 ได้ออกเป็นซีดีรอมบรรจุคลังข้อมูลผู้เรียนภาษาอังกฤษจากผู้เรียนชาติต่างๆ ได้แก่ Bulgarian, Czech, Dutch, Finnish,

[1] ข้อมูลจาก http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm

French, German, Italian, Polish, Russian, Spanish, และ Swedish พร้อมทั้งมีข้อมูลภาษาอังกฤษแบบอังกฤษและแบบอเมริกาเพื่อใช้เป็นคลังข้อมูลเทียบภาษา (LOCNESS[2] - the Louvain Corpus of Native English Essays - British pupils' A level essays: 60,209 คำ, British university students essays: 95,695 คำ, American university students' essays: 168,400 คำ, รวม 324,304 คำ) ปัจจุบันมีข้อมูล 3 ล้านคำเป็นข้อมูลผู้เรียนภาษาอังกฤษจากผู้พูดภาษาต่างๆ 21 ภาษา (มี 21 subcorpora) ได้แก่ Arabic, Bulgarian, Brazilian, Chinese, Czech, Dutch, Finnish, French, German, Greek, Italian, Japanese, Lithuanian, Norwegian, Pakistani, Polish, Portuguese, Romanian, Russian, Slovene, Spanish, South African, Swedish, และ Turkish

ข้อมูลมีการกำกับหมวดคำโดยโปรแกรม ICLE Tagger ที่พัฒนาโดยทีมงาน TOSCA (Tools for Syntactic Corpus Analysis) University of Nijmegen ผู้เรียนที่ให้ข้อมูลกับ ICLE จะต้องกรอกข้อมูลผู้เรียน (ดูภาคผนวกท้ายบท) เพื่อเป็นข้อมูลอ้างอิง งานเขียนที่รวบรวมเป็นงานด้าน argumentative essay ขนาด 500-1,000 คำที่กำหนดให้นักศึกษาเขียนอภิปรายในหัวข้อเรื่องที่กำหนดให้นักเรียนเลือก เช่น "Crime does not pay", "Feminism has done more harm to the cause of women than good", "Pollution : a silent conspiracy" และไม่กำหนดเวลาเขียน นักศึกษาสามารถใช้ทรัพยากรอ้างอิงต่างๆ อย่างพจนานุกรม หนังสือไวยากรณ์ ได้ แต่จะต้องเขียนขึ้นเอง งานเขียนที่เป็นงานสอบหรือที่เป็นงานในชั้นเรียนก็สามารถเก็บได้ แต่ไม่ควรเกิน 25% ของ

[2] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/LOCNESS1.htm

ข้อมูลและต้องมีข้อมูลต่างๆ ชัดเจน ข้อมูล sub-corpus ของ ICLE หรือข้อมูลแต่ละภาษามีขนาดประมาณ 200,000 คำ เก็บจากผู้เรียนประมาณ 300-400 คน

โครงการ USP Learner Corpus (Tagnin 2003) เก็บข้อมูลผู้เรียนภาษาต่างๆ ได้แก่ อังกฤษ เยอรมัน สเปน โดยผู้เรียนทั้งหมดเป็นชาวบราซิล จึงมีลักษณะเป็น multilingual learner corpus ที่ต่างจาก ICLE ที่เป็นคลังข้อมูลผู้เรียนภาษาอังกฤษจากนักเรียนชาติต่างๆ คลังข้อมูลของ USP จึงช่วยตอบคำถามอย่างเช่น อะไรที่เป็นปัญหาร่วมกันไม่ว่าจะเป็นการเรียนภาษาต่างประเทศอะไร หรืออะไรที่เป็นปัญหาเฉพาะของนักเรียนบราซิล ในขณะที่คลังข้อมูล ICLE จะช่วยตอบคำถามอย่างเช่น อะไรที่เป็นปัญหาร่วมกันในการเรียนภาษาอังกฤษไม่ว่าจะเป็นนักเรียนชาติใด

โครงการ Multimedia Adult ESL Learner Corpus (MAELC) (Reder et al. 2003) เป็นโครงการที่จัดทำโดยภาควิชา Applied Linguistics มหาวิทยาลัย Portland State University ร่วมกับ Portland Community College จัดเก็บข้อมูลบักทึกเป็นวีดีโอของการเรียนการสอนภาษาอังกฤษเป็นภาษาต่างประเทศเป็นเวลาห้าปี (2001-2006) มีข้อมูลประมาณ 5,000 ชั่วโมงมีนักเรียนผู้ใหญ่ (adult learner) จำนวนประมาณ 1,000 คน เป็นผู้เรียนระดับต่ำ (low level)

โครงการ Thai English Learner Corpus (TELC) เป็นคลังข้อมูลที่มีการกำกับหมวดคำและรากคำ (part of speech and lemma) ประกอบด้วยงานเขียน 1.5 ล้านคำที่เก็บจากการสอบเข้า (750,000 คำ) และงานที่ให้ทำ (750,000 words) ของนักศึกษาชั้นปีที่ 4 ในมหาวิทยาลัยอัสสัมชัน ซึ่งเก็บตั้งแต่ปีค.ศ. 1997 เคยมีให้ค้นผ่านอินเทอร์เน็ตได้ แต่ปัจจุบันไม่สามารถใช้ได้

โครงการ Longman Learner's Corpus (LLC)[3] เป็นโครงการของเอกชนเพื่อประโยชน์เชิงพาณิชย์และในงานวิจัยสำหรับนักวิจัย มีข้อมูลผู้เรียนภาษาอังกฤษชาติต่างๆ กว่า 10 ล้านคำ ข้อมูลมีการกำกับหมวดคำ และเก็บข้อมูลผู้เรียนภาษาอังกฤษในระดับต่างๆ ซึ่งสามารถดูความสามารถทางภาษาอังกฤษของกลุ่มผู้เรียนต่างๆ เป็นประโยชน์โดยตรงต่อการวิเคราะห์ข้อผิดและความสามารถที่มี ทำให้ผู้ที่ทำพจนานุกรม ผู้ที่ทำหนังสือเรียนต่างๆ กำหนดเนื้อหาและคำอธิบายการใช้ที่เหมาะสมกับผู้เรียนได้

โครงการ Cambridge Learner Corpus (CLC) เป็นโครงการเอกชนเก็บข้อมูลผู้เรียนชาติต่างๆ ทั่วโลก และมีไว้เพื่อใช้ภายในบริษัท มีขนาดกว่า 30 ล้านคำจากนักเรียน 95,000 คนจาก 190 ประเทศและมีภาษาแม่ต่างๆ 130 ภาษา CLC เป็นส่วนหนึ่งของคลังข้อมูล Cambridge International Corpus (CIC) สร้างโดยสำนักพิมพ์ Cambridge University Press และ ESOL (ซึ่งเป็นส่วนหนึ่งของ Local Examination Syndicate มหาวิทยาลัย Cambridge) ข้อมูลภาษาผู้เรียนจึงมาผู้ที่สอบข้อสอบ Cambridge ESOL ทั่วโลก ได้แก่ KET Key English Test, PET Preliminary English Test, FCE First Certificate in English, CAE Certificate in Advanced English, CPE Certificate of Proficiency in English, BEC Business English Certificate, IELTS International English Language Testing System, CELS Certificates in English Language Skills

ผู้ที่ทำพจนานุกรม หนังสือเรียนต่างๆ ให้กับสำนักพิมพ์ Cambridge สามารถเข้าถึงคลังข้อมูลนี้ได้ เพื่อใช้ศึกษาทำความเข้าใจข้อผิดและความสามารถทางภาษาของผู้เรียนภาษาอังกฤษจากที่ต่างๆ ข้อมูลมีการกำกับ

---

[3] http://www.pearsonlongman.com/dictionaries/corpus/learners.html

ข้อผิดด้วยระบบที่สำนักพิมพ์ Cambridge กำหนดขึ้น

โครงการ HKUST Corpus เป็นโครงการขนาดใหญ่แต่มีไว้ใช้ภายในองค์กร มีข้อมูลผู้เรียนกว่า 25 ล้านคำ ไม่มีเว็บไซ้ต์โครงการ ข้อมูลมีการกำกับหมวดคำ (ข้อมูลจาก http://leo.meikai.ac.jp/~tono/lcorpuslist.html)

โครงการ SILS corpus เป็นคลังข้อมูลผู้เรียนภาษาอังกฤษของนักศึกษาใน School of International Liberal Studies ที่มหาวิทยาลัย Waseda นักศึกษามีทั้งที่เป็นชาวญี่ปุ่นและชาวต่างชาติจากที่ต่างๆ บางคนก็เป็นผู้พูดภาษาอังกฤษเป็นภาษาแม่ เป็นข้อมูลที่เก็บเพื่อประโยชน์ในการวิจัยการเรียนการสอนภาษาอังกฤษเป็นภาษาต่างประเทศ (Muehileisen 2006)

โครงการ Polish-English Language Corpus Research and Applications (PELCRA)[4] มีการเก็บคลังข้อมูลภาษาหลายคลังข้อมูล หนึ่งในนั้นคือ Polish Learner English Corpus ซึ่งเป็นคลังข้อมูลสำคัญในโครงการ PELCRA โดยจัดเก็บข้อมูลผู้เรียนระดับต่างๆ และในหลายๆการใช้งาน มีข้อมูลประมาณ 500,00 คำและจะมีการกำกับหมวดคำโดยโปรแกรม CLAWS ของ Lancaster University

โครงการ NICT Japanese Learner Corpus (JLE) Corpus เป็นโครงการเก็บข้อมูลผู้เรียนภาษาอังกษชาวญี่ปุ่น จัดทำโดย National Institute of Information and Communications Technology เป็นข้อมูลภาษาพูดที่เก็บจากตัวอย่าง 1,281 คน มีขนาด 1.2 ล้านคำ ที่ได้จากการสอบ English oral proficiency interview test นับรวมเวลาทั้งหมด 300 ชั่วโมง เดิมใช้ชื่อว่าคลังข้อมูล SST (Standard Spoken Test) ข้อมูลมีการกำกับข้อมูลพื้นฐาน

[4] http://pelcra.ia.uni.lodz.pl/corpora_en.php

(basic tag) ซึ่งจะบอกถึงโครงสร้างตัวบทว่าใครเป็นผู้พูด บอกผลัด บอกการพูดซ้อน เป็นต้น และมีการกำกับข้อผิด (error tag) ในระดับคำและระดับไวยากรณ์ (lexical and grammatical errors) นอกจากนี้ ยังได้จัดทำคลังข้อมูลคู่เทียบที่เก็บข้อมูลพูดของผู้พูดภาษาอังกฤษในบริบทของการทดสอบแบบเดียวกันด้วย จึงมีข้อมูลพร้อมสำหรับศึกษาเปรียบต่างระหว่างการพูดของผู้เรียนชาวญี่ปุ่นกับการพูดของผู้พูดภาษาอังกฤษ และยังจัดทำบทแปลสิ่งที่ผู้เรียนพูดออกเป็นภาษาญี่ปุ่น ทำให้สามารถศึกษาเรื่องการแทรกแซงจากภาษาที่หนึ่งได้ด้วย

โครงการคลังข้อมูลภาษาผู้เรียนชาวไทย (จุฬา-ธรรมศาสตร์) (CU-TU Thai Learner English Corpus) เป็นโครงการภายใต้การสนับสนุนของสภาวิจัยแห่งชาติ ปี 2551-2553 มีวัตถุประสงค์เพื่อสร้างคลังข้อมูลภาษาอังกฤษที่ได้จากงานเขียนของผู้เรียนชาวไทย โดย แยกเป็นสามกลุ่มใหญ่ กลุ่มแรกคือผู้เรียนภาษาอังกฤษระดับมหาวิทยาลัยขั้นพื้นฐานไม่จำกัดคณะ ได้ข้อมูลจากนักศึกษามหาวิทยาลัยธรรมศาสตร์จำนวนกว่า 540,000 คำ จากนิสิตจุฬาลงกรณ์มหาวิทยาลัยจำนวนกว่า 320,000 คำ กลุ่มที่สองเป็นผู้เรียนภาษาอังกฤษระดับมหาวิทยาลัยขั้นสูงโดยเลือกนิสิตเอกภาษาอังกฤษคณะอักษรศาสตร์ ได้ข้อมูลจำนวน 37,900 คำ กลุ่มที่สามเก็บข้อมูลภาษาอังกฤษชาวไทยที่มีความรู้ภาษาอังกฤษดีเขียนภาษาอังกฤษประกอบวิชาชีพ ปัจจุบันเก็บข้อมูลได้ 730,000 คำ โดยจะเปิดให้สืบค้นข้อมูลลักษณะต่างๆได้ผ่านทางอินเทอร์เน็ต

5. ตัวอย่างงานวิจัยที่ใช้คลังข้อมูลภาษาผู้เรียน

Flowerdew (1998) พูดถึงการใช้ประโยชน์จากคลังข้อมูลภาษาผู้เรียนโดยเปรียบเทียบคลังข้อมูลผู้เรียน HKUST กับคลังข้อมูลผู้เชี่ยวชาญ (expert corpus) MCC เพื่อศึกษาสิ่งที่เป็น cause กับ effect marker ว่ามีการใช้ผิด (misuse) ใช้มากเกิน (overuse) ใช้น้อยเกิน (underuse) อย่างไร ซึ่งความ

แตกต่างเหล่านี้มีความสำคัญต่อการจัดการเรียนการสอน ที่คิดกันแต่เดิมว่าจะให้ความสำคัญในการสอนกับคำที่เจ้าของภาษาใช้บ่อยๆ นั้นยังไม่เพียงพอ แต่ควรมองไปถึงการใช้ที่ผู้เรียนแตกต่างจากเจ้าของภาษาด้วย

Altenberg ang Granger (2001) ศึกษาเปรียบเทียบการใช้คำที่มีความถี่การใช้สูง (high frequency word) ของผู้เรียนและเจ้าของภาษา โดยเลือกคำว่า make เป็นกรณีศึกษาเพื่อดูว่าผู้เรียนมีแนวโน้มที่ใช้คำกลุ่มนี้มากเกินไปหรือน้อยเกินไป และการใช้ที่ไม่เหมาะนี้มาจากการถ่ายโอนภาษาหรือไม่ ข้อมูลภาษาผู้เรียนมาจาก ICLE ซึ่งเป็นคนฝรั่งเศสและสวีเดนและเป็นนักศึกษามหาวิทยาลัยปีที่สองและสาม นำมาเปรียบเทียบกับคลังข้อมูลเจ้าของภาษา Louvain Corpus of Native English Essays (LOCNESS) ซึ่งเป็นข้อมูลประเภทเดียวกัน เมื่อวิเคราะห์หน้าที่และความหมายของการใช้ make เป็น 8 ประเภท ได้แก่ produce something, delexical uses, causative uses, earn (money), link verb uses, make it (idiomatic), phrasal/prepositional uses, other conventional uses ก็พบผลว่าผู้เรียนมีการใช้คำ make ในลักษณะ delexical uses น้อยกว่าเจ้าของภาษา และใช้เป็น causative uses แตกต่างกันในกลุ่มผู้เรียน โดยนักเรียนชาวสวีเดนใช้เป็น causative uses มากในขณะที่นักเรียนชาวฝรั่งเศสใช้ในหน้าที่นี้น้อยกว่าเจ้าของภาษา

Sugiura (2002) ศึกษาความรู้เรื่องคำปรากฏร่วมของผู้เรียนภาษาเปรียบเทียบกับเจ้าของภาษา การศึกษาเรื่องคำปรากฏร่วมเป็นหัวข้อหนึ่งที่เป็นที่สนใจในการศึกษาแบบอิงคลังข้อมูลภาษา เพราะคลังข้อมูลจะช่วยให้เห็นถึงรูปแบบการใช้คำปรากฏร่วมที่เจ้าของภาษาใช้ในภาษาของตน ซึ่งความรู้เรื่องคำปรากฏร่วมนี้มักเป็นปัญหาพื้นฐานหนึ่งสำหรับผู้เรียนภาษาต่างประเทศ ที่ผ่านมามีการศึกษาพบว่าผู้เรียนมีการใช้รูปแบบคำปรากฏร่วมที่มากเกินไปและน้อย

เกินไปเมื่อเทียบกับคำปรากฏร่วมที่เจ้าของภาษาใช้ ในงานนี้ Sugiura ตั้งคำถามเรื่องความรู้การใช้คำปรากฏร่วม (collocational knowledge) ของผู้เรียนภาษาอังกฤษชาวญี่ปุ่น โดยศึกษาเปรียบเทียบส่วนที่เป็นวลีตายตัว (fixed expression) ใช้วิธีให้เจ้าของภาษาเขียนประโยคใหม่ (paraphrase) จากข้อมูลที่ได้จากผู้เรียนทีละประโยค จากนั้นใช้โปรแกรม n-gram เพื่อดึงคำที่ติดกันออกมาเป็นข้อมูลของวลีตายตัว วิธีนี้ทำให้ได้ข้อมูลคำปรากฏร่วมที่เทียบเคียงกันได้ชัดเจน ผลก็พบว่าผู้เรียนมีการใช้รูปแบบคำเฉพาะของตนแตกต่างจากเจ้าของภาษา

Neff et al. (2003) ศึกษาการใช้ modal verbs (can, could, may, might and must) และกริยา reporting verbs (suggest, wonder, argue, explain, express, recognise, say, show, and state) ในผู้เรียนภาษาอังกฤษจากพื้นภาษาแม่ต่างๆ ห้าภาษาได้แก่ Spanish, Dutch, Italian, French และ German โดยใช้คลังข้อมูล ICLE พบว่าผู้เรียนมีแนวโน้มใช้ modal verb ทั้งที่มากเกินไป (overuse) หรือใช้น้อยเกินไป (underuse) เมื่อเทียบกับงานเขียนโดยคนอเมริกันที่เป็นเจ้าของภาษา ในส่วนของกริยา reporting verb เจ้าของภาษาจะมีการใช้คำที่หลากหลายมากกว่าและมีหน้าที่ความหมายซับซ้อนกว่า

Abe (2003) ศึกษาข้อผิดจากภาษาพูดและภาษาเขียนของนักศึกษาญี่ปุ่นที่เรียนภาษาอังกฤษ โดยให้ทำข้อทดสอบมาตรฐานที่กำหนดขึ้นซึ่งมีการแบ่งระดับความยากง่ายตามความสามารถผู้ทดสอบจากระดับ 2 ถึง 9 จากนั้นนำข้อมูลมาวิเคราะห์และกำกับข้อผิดต่างๆ ซึ่งพบว่าข้อผิดในภาษาพูดและภาษาเขียนนั้นมีลำดับที่สอดคล้องกัน ส่วนระดับความสามารถผู้เรียนมีผลต่อข้อผิดที่พบ

Lozano and Mendikoetxea (2007) ใช้คลังข้อมูล ICLE เพื่อศึกษาการ

เรียงลำดับคำของประธานและกริยาในภาษาอังกฤษซึ่งมีแบบ inversion ที่วางประธานหลังกริยาและแบบปกติที่วางประธานหน้ากริยาของผู้เรียนชาวสเปนและชาวอิตาเลียน โดยศึกษาทั้งในด้านวากยสัมพันธ์ อรรถศาสตร์ และวัจนปฏิบัติศาสตร์ การศึกษาแต่เดิมมองว่าคำกริยาภาษาอังกฤษมีทั้งแบบที่เป็น unergative เช่น speak และ unaccusative เช่น arrive ในทางวากยสัมพันธ์ subject ของ unergative อยู่ในตำแหน่ง specifier ของ VP ส่วน subject ของ unaccusative อยู่ในตำแหน่ง complement ของ V ในกรณีที่ไม่มีการย้าย subject ไปข้างหน้าก็จะเกิดรูปแบบที่เรียกว่า inversion เช่น There arrived a friend of mine. แต่ในภาษาสเปนและอิตาเลียนการสลับประธานและกริยาจะมีความเป็นอิสระหรือ free inversion มากกว่า ไม่จำกัดเหมือนภาษาอังกฤษ แต่การเลือกสลับประธานและกริยาในภาษาสเปนและอิตาเลียนนั้นมีผู้อธิบายโดยใช้โครงสร้างข้อมูล (information structure) เรื่อง topic, focus หรือ old, new information ซึ่งก็นำไปอธิบาย inversion ที่พบในภาษาอังกฤษได้เช่นกันว่าจะมีการใช้ inversionในกรณีประธานเป็น foucs นอกจากนี้ก็มีเงื่อนไขของการเป็น heavy NP ที่ทำให้เกิด inversion ด้วย แต่การศึกษางานเขียนผู้เรียนภาษาอังกฤษที่ผ่านมามักมองเพียงเงื่อนไขของความเป็น unaccusative สำหรับการทำ inversion ซึ่งผลจากการศึกษาคลังข้อมูล ICLE ของนักเรียนชาวสเปนและชาวอิตาเลียน ก็พบว่าเป็นไปตามเงื่อนไขของการเป็น focus และ heavy NP ตามที่คาด นอกจากนี้เมื่อเทียบกับคลังข้อมูลเจ้าของภาษา LOCNESS ก็พบเงื่อนไขลักษณะเดียวกัน เพียงแต่มความต่างว่าผู้เรียนภาษาสเปนมีการใช้ inversion มากกว่าหรือกล่าวได้ว่ามีการใช้มากเกิน (overuse) เจ้าของภาษา

Waibel (2007) เป็นวิทยานิพนธ์ปริญญาเอกที่ศึกษากริยาเสมือนวลี (phrasal verb) ของผู้เรียนภาษาอังกฤษชาวเยอรมันและชาวอิตาเลียน Phrasal

verbs คือการประกอบกันของคำกริยากับอนุภาควิเศษณ์ (adverbial particle) เช่น make up, put off ซึ่งเป็นปัญหาสำคัญหนึ่งของการเรียนภาษาอังกฤษ เพราะความหมายของ phrasal verb ไม่ใช่ความหมายที่ได้จากการรวมความหมายส่วนย่อยๆ ผู้เรียนภาษาอังกฤษมักมีปัญหาการใช้ phrasal verb และมีการใช้น้อยกว่าเจ้าของภาษาอังกฤษเอง ในงานนี้มีวัตถุประสงค์เพื่อศึกษาว่าผู้เรียนชาวเยอรมันและชาวอิตาเลียนมีรูปแบบการใช้ phrasal verb อย่างไร โดยดึงสิ่งที่จัดว่าเป็น phrasal verb ทั้งหมดที่พบออกมา ไม่ได้ดึงเฉพาะรายการที่กำหนดไว้ และไม่จำกัดว่าเป็น phrasal verb ประเภทไหน (เช่น figurative หรือ literal) phrasal verb ที่มีการใช้น้อย ผู้เรียนมีกลวิธีอะไรที่มาใช้แทนเช่น เลือกใช้คำเดี่ยวที่มีความหมายเหมือนกัน (เช่น put off = postpone) พร้อมทั้งดูว่าภาษาแม่ที่ต่างกันมีอิทธิพลต่อการใช้ phrasal verb อย่างไร เพราะภาษาอิตาเลียนไม่ได้มีคำกริยาที่มีลักษณะแบบ phrasal verb ในภาษาอังกฤษ ในขณะที่ภาษาเยอรมันจะมีคำกริยาที่จัดเป็น phrasal verb คล้ายภาษาอังกฤษ และเมื่อมีการใช้ phrasal verb ผิดก็จะวิเคราะห์ข้อผิดที่พบ

Kamakura (2007) ศึกษา phraseology ของผู้เรียนภาษาอังกฤษชาวญี่ปุ่นโดยการดูกลุ่มคำตั้งแต่ 2 ถึง 4 คำที่มีคำบุพบท in, on, at อยู่ และดูเทียบกับข้อมูลเจ้าของภาษา ข้อมูลผู้เรียนใช้จากคลังข้อมูล Showa Woman's university ที่เก็บจากนักศึกษาปีสามและปีสี่ ส่วนคลังข้อมูลเจ้าของภาษาก็ใช้คลังข้อมูล LOCNESS

6. การกำกับข้อผิด

การกำกับข้อผิดเป็นงานที่ทำกันมาตั้งแต่ในสมัยที่มีความสนใจเรื่องการวิเคราะห์ข้อผิด Dulay, Burt, and Krashen (1982) พูดถึงการจำแนกข้อผิดเป็นสองประเภท ประเภทแรกเป็นข้อมูลภาษา อย่างเช่น การเติมวิภัติปัจจัยผิด การ

ใช้คำผิด ข้อผิดทางไวยากรณ์ ข้อผิดการใช้กริยาช่วย การใช้คำบุพบท เป็นต้น ประเภทที่สองเป็นข้อผิดทางรูป เช่น ตกคำ สะกดผิด สลับที่ เป็นต้น ในปัจจุบัน การกำกับข้อผิดเป็นการกำกับแท็กลงบนคลังข้อมูลภาษาผู้เรียนซึ่งจะมองรวมเป็นระบบเดียว การกำกับข้อผิดไม่ใช่การแก้ไขงานเขียนของนักเรียนให้สละสลวยหรือสมบูรณ์แบบ แต่เป็นการบันทึกว่ามีอะไรที่ผิดไปบ้าง Granger (2003) กล่าวถึงระบบการกำกับข้อผิดที่ดีนั้น ควรมีคุณลักษณะสี่ประการ คือ

1) ทำให้การกำกับข้อมูลเป็นไปแบบคงเส้นคงว่า (consistent) ไม่ควรมีความแตกต่างกันเมื่อเปลี่ยนผู้ที่กำกับข้อผิด การที่จะทำเช่นได้จะต้องมีการจัดทำคู่มือการกำกับข้อผิดที่ละเอียดและชัดเจน

2) ข้อผิดที่กำกับนั้นควรให้ข้อมูลที่เพียงพอแก่ผู้ใช้สำหรับวิเคราะห์ข้อผิดเรื่องต่างๆ ได้ แต่ในขณะเดียวกันก็ต้องไม่มีข้อมูลที่มากเกินไปที่ทำให้จัดการได้ยาก (informative and managebale)

3) มีความยืดหยุ่นเพียงพอที่จะเพิ่มเติมหรือลบแท็กต่างๆ ในระหว่างการกำกับข้อผิด และสามารถเรียกใช้งานได้ง่ายหลังจากที่กำกับข้อผิดเสร็จแล้ว (flexible)

4) ชุดของแท็กต่างๆ ที่ใช้ควรจะใช้ได้กับข้อมูลชุดอื่นๆ หรือภาษาอื่นๆ ด้วย (reusable) เพราะจะทำให้สามารถใช้ระบบเดียวกันในการวิเคราะห์ข้อมูลแบบต่างๆ ได้

ระบบที่ใช้กำกับข้อผิดยังไม่ได้มีข้อตกลงที่เป็นมาตรฐานเดียวกัน แม้ว่าจะมีการมองข้อผิดต่างๆ ในลักษณะที่คล้ายกัน แต่ในรายละเอียดแล้ว แต่ละโครงการวิจัยก็มีระบบการจำแนกข้อผิดที่แตกต่างกัน ในที่นี้จึงจะเสนอระบบการกำกับข้อผิดจากแหล่งต่างๆ มาพอเป็นสังเขป

ระบบการกำกับข้อผิดของ Cambridge Learner Corpus Nicolls

(2003) อธิบายถึงระบบการกำกับข้อผิดใน CLC ว่าได้ใช้รูปแบบการแท็กข้อผิดแบบ <#CODE>wrong word|corrected word</#CODE> ข้อผิดส่วนใหญ่จะใช้ตัวอักษรสองตัวในเครื่องหมายแท็กเพื่อบอกประเภทข้อผิด ตัวอักษรแรกบอกประเภทใหญ่ ตัวอักษรที่สองบอกประเภทย่อย รหัสที่ใช้ในข้อผิดลักษณะนี้เป็นดังนี้

	รหัสตัวที่หนึ่ง		รหัสตัวที่สองบอก word class
F	wrong Form used	A	Pronoun (Anaphoric)
M	something Missing	C	Conjunction (linking word)
R	word or phrase needs Replacing	D	Determiner
U	word or phrase is Unnecessary (i.e. redundant)	J	AdJective
D	word is wrongly Derived	N	Noun
		Q	Quantifier
		T	Preposition
		V	Verb (includes modals)
		Y	Adverb (-lY)

ข้อผิดเกี่ยวกับการใช้เครื่องหมายวรรคตอน (punctuation errors) กำกับด้วยรหัสในตารางข้างล่าง ตัวอย่างเช่น He died <#MP>we|.We</#MP> buried him the next day. หมายความว่าตก full stop ต้องแก้ไขด้วยการแทนที่ we ด้วย . We

MP	punctuation Missing
RP	punctuation needs Replacing
UP	Unnecessary punctuation

ข้อผิดเกี่ยวกับ false friend กำกับด้วยรหัส FF แล้วตามด้วยรหัสบอก word class ส่วนข้อผิดความสอดคล้อง (agreement error) กำกับด้วยรหัส AG แล้วตามด้วยรหัสบอก word class นอกจากนี้ยังมีข้อผิดอื่นๆ ซึ่งกำกับด้วยรหัสตามตารางข้างล่าง

AS	incorrect Argument Structure
CE	Compound Error
CL	CoLlocation error
ID	IDiom error
IN	Incorrect formation of Noun plural
IV	Incorrect Verb inflection
L	inappropriate register (Label)
S	Spelling error
SA	American Spelling
SX	Spelling confusion error
TV	wrong Tense of Verb
W	incorrect Word order
X	incorrect formation of negative

รหัส AS ใช้กรณีที่จำนวน argument ไม่เป็นไปตามที่ควร และไม่สามารถมองเป็นข้อผิดเรื่องตกคำอย่าง *he explained me หรือมีคำเกินอย่าง *he told to me รหัส AS ใช้กรณีเช่นใช้ double object ผิด. *it caused trouble to me จะถูกกำกับเป็น <#AS>it caused trouble to me|it caused me trouble</#AS> รหัส SA (American Spelling) ความจริงไม่ใช่ข้อผิด แต่กำกับไว้เพื่อประโยชน์ให้แน่ใจว่านักเรียนจะรู้ว่าตนเองใช้ภาษาอังกฤษตามแบบไหน

รหัส SX (spelling confusion error) ใช้กำกับข้อมูลคำผิดที่ผู้เรียนสะกดผิดเป็นอีกคำที่คล้ายกัน เช่น to/too, their/there and weather/whether

การกำกับข้อผิดสามารถวางซ้อนกันได้ เช่น I like to <#S>hospitelize|hospitalise</#S> my relatives บอกว่าเป็นการสะกดคำผิด แต่หากดูความหมายที่ต้องการจริงๆ นักเรียนต้องการหมายถึง 'put them up' หรือ 'have them to stay' (show them hospitality) ดังนั้นจึงควรกำกับด้วยแท็ก RV ด้วย <#RV><#S>hospitelize|hospitalise</#S>|put up</#RV>. หมายความว่า hospitelize เป็นการสะกดผิด คำที่ถูกคือ hospitalize แต่จริงๆ ในประโยคนี้ควรแทนด้วยกริยา put up มากกว่า ในการกำกับข้อผิด ผู้วิเคราะห์มักจะพบว่ามีวิธีแก้ที่ผิดได้มากกว่าหนึ่งวิธี เช่น *He said me that ... สามารถแก้เป็น He said to me that หรือ He told me that คำถามคือจะเลือกกำกับข้อผิดอย่างไร หลักการคือพยายามรักษาแบบที่นักเรียนเขียนมากที่สุด ในกรณีนี้ การแก้ argument structure ของคำกริยาจะดูใกล้ของเดิมมากกว่าการแก้คำกริยา จึงกำกับข้อผิดเป็น 'He said <#MT>|to</#MT> me that' มากกว่าที่จะเป็น 'He <#RV>said|told</#RV> me that' อีกทั้งการกำกับแบบแรกทำให้ผู้เรียนเห็นว่าจะใช้กริยา say ให้ถูกอย่างไร ดีกว่าสอนให้เลี่ยงการใช้กริยา say ไปใช้ tell แทน

ระบบการกำกับข้อผิดใน International Corpus of Learner English (ICLE) Granger (2003) กล่าวถึงระบบการกำกับข้อผิดในคลังข้อมูล ICLE ว่ามีพื้นฐานมาจากระบบที่เสนอใน Dagneaux, Denness, and Granger (1998) พร้อมทั้งกล่าวว่าระบบที่ดีจะต้องมีลักษณะดังนี้ ให้ข้อมูลชัดเจนแต่ไม่ละเอียดเกินกว่าจะจัดการได้ สามารถใช้กับข้อมูลภาษาลักษณะอื่นได้ มีความยืดหยุ่น สามารถปรับแก้และเรียกดูได้ง่าย และต้องมีคำอธิบายโดยละเอียดชัดเจนสำหรับ

ให้คนที่กำกับข้อผิดทำงานได้อย่างคงเส้นคงวา

ระบบที่ใช้ใน ICLE ประกอบด้วยการกำกับข้อมูล 3 ระดับ โดเมนของข้อผิด (error domain) ประเภทข้อผิด (error category) และประเภทของคำ (word category) หรือหมวดคำ ทั้งนี้ไม่ได้แยกข้อผิดว่าเป็นข้อผิดจริง (error) หรือเป็นเพียงข้อผิดพลาด (mistake) ให้เป็นเรื่องการวิเคราะห์ของแต่ละคนเอง โดเมนข้อผิดบอกหมวดใหญ่ซึ่งแยกย่อยเป็นประเภทข้อผิดดังแสดงในตารางข้างล่าง ส่วนประเภทของคำกำกับไว้เพื่อสะดวกต่อการวิเคราะห์ เช่น ต้องการหาข้อผิดทั้งหมดที่เกี่ยวกัยคำบุพบท เป็นต้น ประเภทของคำจะกำกับเองด้วยมือเฉพาะส่วนที่ถูกกำกับว่าเป็นข้อผิด ไม่ได้ใช้โปรแกรม POS tagging เพราะไม่แน่ใจความถูกต้องเมื่อใช้กับงานเขียนผู้เรียนภาษา

Error Domains		Error Categories	
<F>	Form	<AGL> <MAJ> <DIA> <HOM> <GRA>	Agglutination Upper/lower case Diacritics Homonymy Other spelling errors
<M>	Morphology	<MDP> <MDS> <MFL> <MFC> <MCO>	Derivation-prefixation Derivation-suffixation Inflection Inflection-confusion Compounding
<G>	Grammar	<CLA> <AUX> <GEN> <MOD> <NBR> <PER> <TPS> <VOI> <EUF>	Class Auxiliary Gender Mode Number Person Tense Voice Euphony

7. บทสรุป

คลังข้อมูลภาษาผู้เรียนเป็นทรัพยากรที่มีความสำคัญต่อการศึกษาวิจัยเรื่องการเรียนการสอนภาษาต่างประเทศเป็นอย่างมากในปัจจุบัน แม้ว่าการศึกษาเปรียบต่างภาษาอังกฤษและภาษาไทยจะมีผู้ทำมาเป็นเวลานานแล้ว แต่ก็เป็นการเปรียบต่างโดยไม่ได้อาศัยข้อมูลภาษาผู้เรียนเป็นหลัก การศึกษาข้อผิดที่ผ่านมาก็ทำให้เห็นเฉพาะข้อผิดแต่ไม่ได้เห็นถึงความสามารถของผู้เรียนว่ามีความเหมือนหรือต่างจากเจ้าของภาษามากน้อยเพียงใด อีกทั้งการไม่มีข้อมูลกลางที่สามารถใช้ร่วมกันได้ ก็เป็นอุปสรรคต่อการศึกษาและหาข้อสรุปร่วมกัน การพัฒนาคลังข้อมูลผู้เรียนภาษาอังกฤษชาวไทยในครั้งนี้จึงมีความจำเป็นและจะเป็นประโยชน์ต่อการศึกษาวิจัยการเรียนการสอนภาษาอังกฤษในประเทศไทย

ภาคผนวก ICLE Learner Profile

```
LEARNER PROFILE
================================================================
Text code : (do not fill in)
Essay :
Title :
Approximate length required : -500 words 0  +500 words  0
Conditions : timed 0  untimed 0
Examination : yes 0   no 0
Reference tools : yes 0  no 0
What reference tools ?
            Bilingual dictionary :
            English monolingual dictionary :
            Grammar :
            Other(s) :
================================================================
Surname :                      First names :
Age :        Male  0    Female  0
Nationality :
Native language :
Father's mother tongue :
Mother's mother tongue :
Language(s) spoken at home : (if more than one, please give the
average % use of each)
Education :
Primary school - medium of instruction :
```

```
Secondary school - medium of instruction :
Current studies :
Current year of study :
Institution :
Medium of instruction :
            English only 0
            Other language(s) (specify) 0
            Both    0
================================================================
Years of English at school :
Years of English at university :
Stay in an English-speaking country :
Where ?
When ?
How long ?
================================================================
Other foreign languages in decreasing order of proficiency :
================================================================
     I hereby give permission for my essay to be used for
research purposes.
Date :         Signature :
```

บทที่ 8 การออกแบบและการสร้างคลังข้อมูลภาษา

1. ข้อควรพิจารณา

ในบทนี้จะกล่าวถึงการออกแบบและการจัดสร้างคลังข้อมูลภาษาซึ่งเป็นเรื่องที่สำคัญมาก ทั้งนี้เพราะการที่จะสามารถใช้ประโยชน์จากคลังข้อมูลภาษาได้ตามที่ต้องการ คลังข้อมูลนั้นจะต้องได้รับการออกแบบโครงสร้างมาให้ได้ข้อมูลครอบคลุมอย่างที่ต้องการ Kennedy (1998) ได้กล่างถึงปัจจัยที่ควรคำนึงในการสร้างคลังข้อมูลภาษาไว้หลายประการ ซึ่งเป็นคำถามที่ผู้สร้างคลังข้อมูลภาษาจะต้องตัดสินใจ ดังนี้

1.1 วัตถุประสงค์

ผู้สร้างจะต้องรู้จุดมุ่งหมายที่แน่นอนว่าต้องการสร้างคลังข้อมูลขึ้นมาเพื่อประโยชน์อะไร ใครคือคนที่จะใช้ประโยชน์จากคลังข้อมูลภาษาที่สร้าง คลังข้อมูลที่สร้างจะใช้เป็นคลังข้อมูลแบบทั่วไป (general corpus) หรือเป็นคลังข้อมูลเฉพาะทาง (specialized corpus) ที่สร้างขึ้นเพื่อวัตุประสงค์เฉพาะเจาะจง เช่น ใช้เพื่อศึกษาพัฒนาการการเรียนรู้ภาษาของเด็ก ใช้เพื่อฝึกสอนระบบการประมวลผลภาษาธรรมชาติกับเอกสารที่เกี่ยวข้องกับเรื่องทางวิชาการ ใช้เพื่อสอนภาษาให้คนต่างชาติ หรือใช้เพื่อศึกษารูปแบบการเขียนของนักเขียนบางคน เป็นต้น วัตถุประสงค์เหล่านี้จะเป็นตัวกำหนดขอบเขตของข้อมูลภาษาที่เกี่ยวข้อง

1.2 ลักษณะของคลังข้อมูลเป็นแบบสถิตย์หรือแบบพลวัต

คลังข้อมูลที่ต้องการจะมีลักษณะเป็นคลังข้อมูลแบบสถิตย์ (static) หรือแบบพลวัต (dynamic) แบบสถิตย์หมายถึงมีขนาดแน่นอนตามที่กำหนด คลังข้อมูลภาษาในยุคแรกจะมีลักษณะเป็นแบบนี้ เช่น คลังข้อมูล SEU หรือแม้แต่คลังข้อมูล BNC ก็จัดอยู่ในประเภทนี้ คลังข้อมูลแบบนี้เป็นเสมือนภาพของภาษาที่ถูกจับหยุดนิ่งในช่วงเวลาหนึ่งๆ เนื่องจากการมีขนาดที่จำกัด คลังข้อมูลแบบนี้จึงเลือกใช้วิธีสุ่มข้อมูลภาษาบางส่วนจากเอกสารแทนที่จะเก็บข้อมูลเอกสารทั้งฉบับ ทั้งนี้เพื่อให้ได้เอกสารหลากหลายประเภทมากขึ้น เช่น LOB และ Brown ได้เลือกเอกสาร 15 ประเภท (genres) ซึ่งประกอบด้วยตัวบทประมาณ 500 ตัวบท ในแต่ละตัวบทมีคำ 2,000 คำซึ่งตัดตอนมาจากเอกสารฉบับเต็ม ปัญหาของคลังข้อมูลแบบสุ่มตัวอย่าง (sample corpora) คือ คำบางคำอาจพบได้น้อยหรือไม่พบเลยในตัวบทขนาดเพียง 2,000 คำนี้ และถ้าต้องการศึกษาเรื่องที่จำเป็นต้องใช้ตัวบทที่มีความยาวมาก เช่น การศึกษาในระดับปริเฉจก็จะทำไม่ได้

ส่วนคลังข้อมูลที่เป็นแบบพลวัตหรือที่ Sinclair เรียกว่า monitor corpus เป็นคลังข้อมูลที่ไม่ได้กำหนดขนาดไว้ตายตัว สามารถขยายใหญ่ขึ้นได้เรื่อยๆตามข้อมูลที่ได้มาใหม่ๆ ขนาดและองค์ประกอบภายในคลังข้อมูลจึงเปลี่ยนแปลงได้ตลอด คลังข้อมูลแบบนี้แทนภาพของภาษาที่ไม่หยุดนิ่ง การเปลี่ยนแปลงใดๆจะถูกเก็บบันทึกไว้ แต่คลังข้อมูลแบบนี้มักจะมีลักษณะที่ไม่สมดุล (balance) อยู่ด้วย และมักเป็นการเก็บข้อมูลแบบตามสะดวก (opportunistic) Sinclair (1991) เสนอว่า ในปัจจุบัน คอมพิวเตอร์สามารถเก็บข้อมูลได้มาก สามารถประมวลผลเอกสารใหญ่ๆได้ จึงควรสนใจสร้าง monitor corpus มากกว่า เรากำลังอยู่ในยุคที่ข้อจำกัดด้านเนื้อที่เก็บข้อมูลนั้นหมดไป เอกสารอิเล็กทรอนิกส์ก็

มีอยู่แพร่หลายมากมาย สิ่งที่เราควรทำคือสร้างโปรแกรมเพื่อกรองแยกแยะตัวบทที่รับเข้ามาเพื่อจัดกลุ่มจัดประเภทของตัวบทนั้นๆ นักวิจัยคนใดต้องการตัวบทประเภทใดก็สามารถดึงตัวบทเฉพาะประเภทที่ต้องการได้ หรือหากจำเป็นต้องการสุ่มดึงข้อมูลบางส่วนของตัวบทก็สามารถทำได้ในภายหลังเช่นกัน ถ้าหากเราสามารถทำได้ดังนี้ คลังข้อมูลแบบพลวัตจะเป็นตัวแทนของภาพภาษาที่เกิดขึ้นในช่วงใดช่วงหนึ่ง (represent state of the language) และเมื่อเก็บข้อมูลภาษาไปเรื่อยๆ เป็นเวลานานพอก็จะมีมิติของเวลาเข้ามา ทำให้เราสามารถเปรียบเทียบความแตกต่างของภาษาในแต่ละช่วงเวลาได้

1.3 การเป็นตัวแทนภาษาและความสมดุล

ปัญหาเรื่องการสร้างคลังข้อมูลที่สามารถเป็นตัวแทนภาษาและมีความสมดุล (representativeness and balance) ก็เป็นเรื่องสำคัญเรื่องหนึ่งที่ต้องคำนึงถึง คำถามที่จะต้องถามต่อก็คือ คลังข้อมูลภาษาที่ต้องการสร้างเป็นตัวแทนของภาษาอะไรหรือภาษาประเภทไหน สิ่งที่คนส่วนมากต้องการคือให้เป็นตัวแทนของภาษาที่เป็นภาษาประจำชาติ และคำถามที่สำคัญอีกคำถามหนึ่ง คือ คลังข้อมูลที่สมดุลเป็นอย่างไร จะเลือกให้ความสำคัญกับเอกสารประเภทใดมากกว่า จะให้ความสำคัญกับภาษาเขียนหรือภาษาพูดมากกว่ากัน คนทั่วไปอาจใช้ภาษาพูดมากกว่าภาษาเขียนในชีวิตประจำวัน แต่ภาษาเขียนที่ปรากฏในหนังสือพิมพ์ก็มีผู้อ่านนับล้านคน ส่วนภาษาพูดที่เกิดขึ้นก็อาจเกิดขึ้นระหว่างคนสองคน หรืออาจเป็นคำปราศรัยที่พูดกับคนนับล้านคน คลังข้อมูล BNC ให้ความสำคัญกับภาษาเขียนมากกว่าโดยให้น้ำหนักของภาษาเขียนเป็น 90% ส่วนคลังข้อมูล ICE ให้น้ำหนักภาษาพูดมากกว่าโดยกำหนดให้มีภาษาพูด 60% หรือแม้แต่ในภาษาเขียนเอง การเลือกให้น้ำหนักกับตัวบทประเภทไหนมากกว่าประเภทไหน ก็เป็นเรื่องที่ต้องคำนึงจากปัจจัยหลายๆด้าน การมีสมดุลไม่ได้

หมายถึงการที่จะต้องให้มีจำนวนตัวบทในแต่ละประเภทในปริมาณเท่าๆกัน Sinclair (1991) เสนอว่าในการพิจารณาเลือกเอกสารภาษาเขียนอย่างน้อยควรแยกความแตกต่างของสิ่งเหล่านี้ คือ นิยาย/ไม่ใช่นิยาย หนังสือ/วารสาร/หนังสือพิมพ์ เอกสารทางการ/ไม่เป็นทางการ Summers (1991) ได้เสนอให้พยายามเลือกตัวบทให้หลากหลายมากที่สุด โดยอาจใช้เกณฑ์ต่างๆในการพิจารณาคัดเลือกตัวบท เช่น เลือกตัวบทที่มีอิทธิพลหรือความสำคัญ เลือกตัวบทที่แพร่หลายอยู่ในปัจจุบัน เลือกตัวบทที่เป็นที่นิยม เลือกตามสัดส่วนลักษณะการอ่าน เป็นต้น นอกจากนี้ ยังอาจต้องพิจารณาในของเรื่องเชื้อชาติ สัญชาติด้วย เช่น เราจะนับรวมภาษาของคนที่เป็นผู้อพยพหรือไม่ หากนับเฉพาะผู้พูดที่เป็นเจ้าของภาษา (native speaker) ใครที่จะจัดว่าเป็นผู้พูดที่เป็นเจ้าของภาษาที่แท้จริง จะใช้เกณฑ์ในการเกิด หรือใช้เกณฑ์ว่าเป็นคนที่มาอาศัยอยู่ในประเทศนั้นนานแล้ว จะนับรวมคนที่ย้ายตามพ่อแม่ไปอยู่ต่างประเทศตั้งแต่ยังเล็กไหม ส่วนคนที่ไม่ได้พูดด้วยสำเนียงภาษาแท้ แต่เป็นบุคคลสำคัญ เช่น Henry Kissinger ซึ่งคำพูดของเขาถือได้ว่าเป็นที่แพร่หลายรู้จักของคนทั่วไป จะรวมไว้ในคลังข้อมูลไหม เหล่านี้ล้วนเป็นคำถามที่ผู้จัดสร้างคลังข้อมูลจะต้องพิจารณาตัดสินใจให้ชัดในการออกแบบคลังข้อมูลเพื่อจะได้เลือกข้อมูลได้ง่ายขึ้น

1.4 ขนาดของคลังข้อมูลภาษา

ขนาดของคลังข้อมูลภาษาก็เป็นอีกเรื่องที่ต้องคำนึงถึง ไม่ว่าขนาดของคลังข้อมูลจะใหญ่เพียงใดก็ตาม ในความเป็นจริงก็เป็นเพียงส่วนเล็กน้อยมากเมื่อเทียบกับจำนวนของภาษาที่เกิดขึ้นทั้งหมด (หากลองนึกถึงปริมาณประโยคที่คนไทยทั้งหมดพูดหรือเขียนในแต่ละวัน) คลังข้อมูลภาษาหนึ่งล้านคำในยุคแรกของการจัดเก็บคลังข้อมูลภาษานั้น จึงจัดว่าน้อยมากเมื่อเทียบกับจำนวนคำหรือประโยคที่เกิดขึ้นจริง แต่ในยุคนั้น ก็ต้องถือว่าเหมาะสม เพราะประสิทธิภาพของ

คอมพิวเตอร์ในขณะนั้นยังต่ำอยู่ หากจะค้นหาคำบางคำ คอมพิวเตอร์อาจต้องทำงานนานหลายชั่วโมงทีเดียว ในขณะที่ปัจจุบัน ประสิทธิภาพของคอมพิวเตอร์สูงขึ้นมาก สามารถค้นหาคำที่ต้องการได้ภายในไม่กี่วินาที ดังนั้น คลังข้อมูลภาษาที่เหมาะสมสำหรับการศึกษาเรื่องภาษาจึงควรมีขนาดใหญ่มากเท่าที่จะสามารถทำได้เพื่อให้ได้ภาพของภาษามากที่สุดเท่าที่จะเป็นไปได้ Sinclair (1991) แนะว่าขนาดของคลังข้อมูลภาษาที่เป็นคลังข้อมูลแบบทั่วไป (general corpus) ควรมีขนาดอย่างน้อย 10-20 ล้านคำ

อีกประเด็นหนึ่งที่ต้องพิจารณา คือจำนวนตัวอย่างคำ (token) ที่ถือว่าเพียงพอที่เราจะใช้เป็นหลักฐานในการอธิบายคำนั้นๆ Kennedy (1998) กล่าวว่า โดยปกติแล้ว ในคลังข้อมูลหนึ่งล้านคำ จะมีคำ 40 - 50% ที่ปรากฏเพียงแค่ครั้งเดียว และในกรณีของคำที่มีหลายความหมาย (polysemous) อย่างน้อยครึ่งหนึ่งของความหมายที่เป็นไปได้ของคำนั้นมีปรากฏใช้เพียงครั้งเดียว เช่น คำว่า *break* ปรากฏ 8,267 ครั้งใน Longman/Lancaster Corpus แต่ *break* ในความหมายของข่าวด่วน (*news breaking)* หรือ *break* ในการแข่งขันเทนนิสกลับไม่ปรากฏมากพอที่จะใช้ในการวิเคราะห์ แต่คำบางคำกลับมีปรากฏมากมาย เช่น คำว่า at อาจปรากฏถึง 5,500 ครั้งในคลังข้อมูลขนาดหนึ่งล้านคำ ซึ่งการที่มีข้อมูลมากเกินไปนั้น ก็เป็นการยากที่จะวิเคราะห์ข้อมูลทั้งหมดที่ปรากฏได้ ต้องเลือกสุ่มตัวอย่างจากผลที่ได้ แต่เราจะใช้หลักอะไรในการสุ่มตัวอย่างเหล่านี้ และขณะนี้ก็ยังขาดโปรแกรมที่จะช่วยนักวิเคราะห์จัดการกับการเลือกข้อมูลมากๆที่ได้

Sinclair แนะว่าคลังข้อมูลภาษาควรจะมีขนาดใหญ่มากที่สุดเท่าที่จะสามารถทำได้ ทั้งนี้เพราะโดยธรรมชาติของตัวภาษาเอง ที่ทำให้คำที่เป็นคำไวยากรณ์ (function word) จะมีการปรากฏใช้มากกว่าคำที่เป็นคำเนื้อหา

(content word) ทำให้เกิดลักษณะที่ไม่สมดุลของการกระจายของคำเป็นอย่างมาก ถ้าต้องการศึกษาคำที่เป็นเนื้อหา คลังข้อมูลต้องมีขนาดใหญ่เพียงพอเพื่อที่จะให้ได้ตัวอย่างของการใช้คำที่ต้องการศึกษาเป็นจำนวนที่มากพอ นอกจากนี้ ถ้าเราต้องการศึกษามากกว่ารูปคำ (form) เช่น ศึกษาการใช้ (usage) ในแบบต่างๆ ก็จะพบปัญหาของปริมาณความไม่สมดุลในการใช้แต่ละแบบด้วย เช่น ในคำที่มีหลายความหมาย การใช้ในความหมายแรกจะปรากฏเป็นจำนวนมากกว่าร้อยละ 90 ของคำนั้น ในขณะที่การใช้ในอีกความหมายหนึ่งอาจพบเพียงครั้งเดียว เป็นต้น หากเราต้องการศึกษาการใช้คำนั้นในความหมายหลัง เราจึงต้องอาศัยคลังข้อมูลภาษาที่ใหญ่มากพอที่จะพบการใช้คำในความหมายนั้นมากกว่าหนึ่งครั้ง สรุปก็คือ ยิ่งคลังข้อมูลภาษามีขนาดใหญ่มากเพียงใดก็ยิ่งเป็นประโยชน์มากเท่านั้น อย่างไรก็ตาม ข้อสรุปนี้ไม่ได้หมายความว่า งานทุกงานจำเป็นต้องอาศัยคลังข้อมูลขนาดใหญ่ งานบางงาน อย่างเช่น การศึกษา prosody คลังข้อมูลภาษาพูด 100,000 คำก็อาจจะพอเพียงถ้าเป็นการเก็บคำพูดแบบต่อเนื่อง หรือถ้าศึกษาเรื่องวจีวิภาค (morphology) คลังข้อมูลภาษาขนาด 500,000 คำก็ควรจะเพียงพอสำหรับใช้ศึกษา (Kennedy 1998: 68) นอกจากนี้ ขนาดของคลังข้อมูลเพียงอย่างเดียวไม่ได้บอกความเป็นตัวแทนของภาษา คลังข้อมูลใหญ่ขนาด 300 ล้านคำไม่ได้เป็นหลักประกันว่าคลังข้อมูลนั้นจะถือเป็นตัวแทนของภาษานั้นได้ ความเป็นตัวแทนของภาษาขึ้นอยู่กับการออกแบบและเก็บตัวอย่างให้ครอบคลุมภาษาในแบบต่างๆตามที่ออกแบบไว้

นอกจากเรื่องขนาดของคลังข้อมูลแล้ว ยังมีเรื่องขนาดของตัวอย่าง (sampling) ว่าควรจะเป็นเท่าไร คลังข้อมูล LOB และ Brown Corpus ใช้ปริมาณ 2,000 คำเป็นเกณฑ์ในคัดแต่ละตัวอย่าง แต่อย่างไรก็ตามตัวเอกสารจริงๆจะมีขนาดต่างกัน แต่ละส่วนของเอกสารมีความสำคัญแตกต่างกัน การ

เลือกตัวอย่างจากตอนต้น ตอนกลาง หรือตอนท้าย ย่อมให้ภาพที่ต่างกันของเอกสารนั้นๆ ถ้าจะเลือกเก็บเอกสารทั้งฉบับ เราก็ไม่ต้องวิตกกับปัญหาว่าจะขาดส่วนใดส่วนหนึ่งไป แต่ก็มีข้อเสียเรื่องความสิ้นเปลือง แทนที่จะเก็บจำนวนเอกสารประเภทต่างๆได้มาก ก็จำเป็นต้องลดจำนวนเอกสารที่เก็บลง (เมื่อคิดจากค่าใช้จ่ายในการสร้างคลังข้อมูลภาษาในขนาดที่เท่ากัน) หรือการคัดเลือกคำเพียง 2,000 คำที่ทำใน Brown Corpus จะทำให้ภาพของการเป็นตัวแทนเอกสารในประเภทนั้นๆหายไปหรือไม่ Biber (1990) พยายามตอบคำถามนี้ โดยศึกษาลักษณะของตัวบทประเภทต่างๆที่ปรากฏใน Brown Corpus ว่าแตกต่างกันหรือไม่ โดยกำหนดจากการพิจารณาลักษณะทางภาษาศาสตร์สิบลักษณะ แล้วคำนวณหาความแตกต่างในเชิงสถิติ ซึ่งเขาได้ข้อสรุปว่าขนาดของตัวอย่าง 2,000 – 5,000 คำนั้นเพียงพอสำหรับจำแนกตัวบทประเภทต่างๆ จากกันได้ หากจำเป็นต้องสุ่มข้อความจากเอกสารเพื่อสร้างคลังข้อมูลภาษา ปริมาณที่สุ่มจึงควรมีปริมาณมากพอสมควรเท่าที่จะเป็นไปได้ และต้องไม่น้อยกว่า 2,000 คำนี้

1.5 การป้อนข้อมูลเข้า

การป้อนข้อมูลเข้าสามารถทำได้หลายวิธี วิธีพื้นฐานที่สุดคือพิมพ์เข้าคอมพิวเตอร์โดยตรง แต่วิธีนี้ต้องเสียเวลามาก วิธีที่เร็วกว่าคือ การใช้เครื่องสแกนเนอร์สแกนเอกสารที่ต้องการเป็นภาพแล้วใช้โปรแกรมรู้จำตัวอักษร (optical character recognition OCR) เพื่อแปลงจากภาพนั้นเป็นตัวเอกสารอิเล็กทรอนิกส์เก็บในคอมพิวเตอร์อีกที ซึ่งวิธีนี้จะได้ผลดีถ้าโปรแกรม OCR มีประสิทธิภาพสูง สามารถรับรู้ (recognize) ตัวอักษรได้ถูกต้อง เช่น โปรแกรม OCR ของภาษาอังกฤษอย่างเช่น โปรแกรม Omni Page สามารถแปลงเอกสารภาษาอังกฤษได้อย่างถูกต้องแม่นยำสูง ส่วนอีกวิธีที่เร็วกว่านั้น คือรวบรวมจากเอกสารอิเล็กทรอนิกส์ที่มีเก็บไว้อยู่แล้วในคอมพิวเตอร์ เช่น เอกสารที่มีอยู่ตาม

สำนักพิมพ์ต่างๆ เอกสารที่เผยแพร่อยู่บนอินเทอร์เน็ต ในการจัดสร้างคลังข้อมูลภาษา คงต้องพิจารณาความเหมาะสมของแต่ละวิธีซึ่งก็มีข้อดีข้อเสียของตัวเอง ให้เลือกเอาวิธีที่เหมาะสมหรือจะใช้ทั้งสามวิธีผสมกันก็ได้ ถ้าจะใช้โปรแกรม OCR กับเครื่องสแกนเนอร์ก็ต้องคำนวณค่าใช้จ่าย ราคาของเครื่องสแกนเนอร์ เวลาที่ใช้ ประสิทธิภาพของโปรแกรม และคำนึงถึงลักษณะตัวอักษรในเอกสารต้นฉบับ ถ้าเป็นฟอนท์ที่แปลกจากปกติ โอกาสที่โปรแกรม OCR จะรู้จักตัวอักษรในฟอนท์นั้นก็จะลดลง

2. ขั้นตอนในการจัดสร้างคลังข้อมูลภาษา

Kennedy (1998) ได้สรุปถึงขั้นตอนที่สำคัญของการจัดสร้างคลังข้อมูลเอาไว้ 5 ขั้นตอน ดังนี้

2.1 การออกแบบโครงสร้าง

การออกแบบโครงสร้างคลังข้อมูลภาษาถือเป็นขั้นตอนแรกของการจัดสร้างคลังข้อมูลภาษา โดยการพิจารณาเรื่องต่างๆที่เกี่ยวข้อง เช่น ประเภทของคลังข้อมูล ขนาด และองค์ประกอบภายใน และควรพิจารณาถึงประโยชน์ในการศึกษาเปรียบเทียบด้วย เพราะคลังข้อมูลภาษาที่สร้างอาจจะนำไปใช้ศึกษาเปรียบเทียบกับคลังข้อมูลภาษาอื่นๆได้ด้วย ปัจจัยหลักที่จะเป็นตัวกำหนดโครงสร้างของคลังข้อมูล ก็คือวัตถุประสงค์ของการใช้งาน ผู้จัดสร้างจะต้องรู้ชัดเจนว่าจะใช้คลังข้อมูลนั้นๆ เพื่อประโยชน์อะไรบ้าง งานบางอย่างก็อาจจะไม่จำเป็นต้องอาศัยคลังข้อมูลขนาดใหญ่ ภาษาที่ต้องการเป็นภาษาพูดหรือภาษาเขียน โครงสร้างของคลังข้อมูลควรประกอบด้วยตัวบทประเภทใดบ้าง นอกจากนั้นยังต้องพิจารณาถึงสัดส่วนที่ควรจะเป็นด้วย ซึ่งการกำหนดสัดส่วนนั้นไม่ว่าอย่างไรก็เป็นการกำหนดแบบไม่มีหลักการ (arbitrary) หรืออย่างดีก็เป็น

การเลือกแบบ informed judgment คือบอกไว้ชัดเจนว่าเลือกอะไรเป็นสัดส่วนเท่าใด แต่ในความเป็นจริง ไม่มีใครตอบได้ว่าสัดส่วนที่แท้จริงของภาษาพูดหรือภาษาเขียนควรเป็นอย่างไร และควรประกอบด้วยตัวบทประเภทใดบ้างในสัดส่วนเท่าใด หลังจากที่ตัดสินใจแบ่งประเภทตัวบทต่างๆ แล้ว ควรจะทำโครงการนำร่อง (pilot study) เพื่อศึกษาดูว่ามีข้อมูลเอกสารในกลุ่มที่ต้องการหรือไม่ นอกจากนี้ ต้องพิจารณาเรื่องการคัดเลือกเอกสารว่าต้องการเก็บเอกสารทั้งฉบับหรือเป็นการเก็บบางส่วน หากเอกสารที่ได้มามีจำนวนคำไม่ครบตามที่กำหนดจะทำอย่างไร ส่วนข้อมูลที่บอกตัวแปรอื่นๆนอกเหนือจากด้านภาษา (extralinguistic variables) ก็ควรจะเก็บไว้ด้วย เช่น แหล่งที่มาของเอกสาร สื่อของเอกสาร ผู้ที่ร่วมในเหตุการณ์ ประเภท (genre) สไตล์ วันที่ สถานที่หรือสถานการณ์ (setting) อายุและเพศของผู้รับสาร เป็นต้น

ในการจัดเก็บข้อมูลเอกสารจริงๆ ผู้จัดทำจะพบปัญหาต่างๆที่ต้องตัดสินใจ ตัวอย่างเช่น ผู้จัดทำอาจมองว่าตัวบทประเภทข่าวนั้นจะเก็บได้จากหนังสือพิมพ์ แต่หนังสือพิมพ์นั้นก็ไม่ได้มีจำกัดเฉพาะสิ่งที่เป็นข่าว แต่มีเรื่องอื่นๆนอกเหนือจากข่าวอยู่ด้วย จะเลือกเฉพาะสิ่งที่เป็นข่าวอย่างเดียวหรือจะเลือกให้คลุมภาพทั้งหมดที่เป็นจริงของหนังสือพิมพ์นั้น นอกจากนี้หนังสือพิมพ์หรือวารสารในแต่ละช่วงของปีก็อาจมีน้ำหนักของตัวบทประเภทต่างๆ ไม่เท่ากัน เช่น ในช่วงเทศกาลสำคัญ หนังสือพิมพ์ วารสารก็จะลงเรื่องที่เกี่ยวข้องกับเทศกาลนั้นเป็นจำนวนมาก ซึ่งอาจจะหลีกเลี่ยงปัญหานี้โดยใช้วิธีกระจายการเก็บข้อมูลจากช่วงวันหรือสัปดาห์ต่างๆ เป็นต้น

2.2 วางแผนการเก็บข้อมูล

คลังข้อมูลภาษาอังกฤษขนาด 1 ล้านคำจะกินเนื้อที่ประมาณ 8-10 ล้านไบต์ (Mbytes) ถ้ามีการกำกับข้อมูลทางไวยากรณ์ด้วยขนาดก็จะเพิ่มขึ้นอีก 3-5

ล้านไบต์ ถ้ากำกับข้อมูลการแจงส่วนประโยคด้วย คือมีการกำกับหมวดคำและโครงสร้างวากยวัมพันธ์ด้วย ขนาดของคลังข้อมูลก็จะเพิ่มอีก 30 ล้านไบต์ ในปัจจุบัน เรื่องเนื้อที่สำหรับจัดเก็บข้อมูลไม่น่าจะเป็นปัญหาอีกต่อไป สิ่งที่ต้องพิจารณา คือการเข้าถึงข้อมูลได้ง่าย สะดวกต่อการใช้งานจริง ซึ่งอาจจำเป็นต้องเก็บข้อมูลเกี่ยวกับตัวบทนั้น เช่น ใครเป็นผู้แต่ง เป็นตัวบทประเภทไหน เกิดขึ้นเมื่อใด ผู้รับสารเป็นใคร เป็นต้น หรือหากเก็บข้อมูลภาษาพูดก็ควรจะมีกลไกที่ให้ผู้ใช้สามารถดูคำถอดเสียง (transcription) พร้อมกับฟังเสียงที่อัดไว้ด้วยได้ และควรจะมีการจัดระบบหมวดหมู่ (catalog) ของแฟ้มข้อมูลทั้งหมด และในระหว่างการจัดสร้างคลังข้อมูลภาษา ก็ควรต้องมีการเก็บสำรองข้อมูลอย่างสม่ำเสมอด้วยเพื่อป้องกันข้อมูลสูญหาย

2.3 ขออนุญาตเจ้าของลิขสิทธิ์

เพื่อหลีกเลี่ยงปัญหาทางกฎหมาย ผู้สร้างคลังข้อมูลภาษาจะต้องขออนุญาตจากเจ้าของลิขสิทธิ์ของตัวบททุกตัวบทที่นำมาใช้ในการจัดทำคลังข้อมูลภาษา และอาจจะต้องติดต่อกับบุคคลที่เป็นผู้พูดในเทปบันทึกเสียงเพื่อขออนุญาตในการรวบรวมไว้ในคลังข้อมูลภาษา เพื่อที่จะได้ไม่มีปัญหาในการละเมิดสิทธิส่วนบุคคล ในบางกรณี เ จ้าของลิขสิทธิ์อาจเรียกร้องค่าธรรมเนียม การสืบค้นว่าใครเป็นเจ้าของลิขสิทธิ์งานนั้นอย่างแท้จริงก็เป็นอีกปัญหาหนึ่งเพราะอาจมีการเปลี่ยนโอนเจ้าของหรือบริษัทได้ โดยปกติ หากขอไปใช้เพื่อประโยชน์ในงานวิจัยมักจะไม่เป็นปัญหา แต่ก็ใช้เวลาพอสมควรเพราะเจ้าของลิขสิทธิ์จะต้องแน่ใจว่าผู้ขอจะไม่นำเอาเอกสารที่ได้ไปใช้ประโยชน์ทางการค้าอื่นๆ

สำหรับประเทศไทย กฎหมายลิขสิทธิ์ฉบับปัจจุบันที่บังคับใช้ คือ พระราชบัญญัติลิขสิทธิ์พ.ศ. ๒๕๓๗ ซึ่งกำหนดให้ลิขสิทธิ์ตามพระราชบัญญัตินี้ให้มีอยู่ตลอดอายุของผู้สร้างสรรค์และมีอยู่ต่อไปอีกเป็นเวลาห้าสิบปี นับแต่ผู้

สร้างสรรค์ถึงแก่ความตายในกรณีที่มีผู้สร้างสรรค์ร่วมลิขสิทธิ์ในงานดังกล่าวให้มีอยู่ตลอดอายุของผู้สร้างสรรค์ร่วม และมีอยู่ต่อไปอีกเป็นเวลาห้าสิบปีนับแต่ผู้สร้างสรรค์ร่วมคนสุดท้ายถึงแก่ความตาย ส่วนข้อยกเว้นของการละเมิดลิขสิทธิ์ในส่วนที่เกี่ยวข้อง มีดังนี้

มาตรา ๓๒ การกระทำแก่งานอันมีลิขสิทธิ์ของบุคคลอื่นตามพระราชบัญญัตินี้ หากไม่ขัดต่อการแสวงหาประโยชน์จากงานอันมีลิขสิทธิ์ตามปกติของเจ้าของลิขสิทธิ์ และไม่กระทบกระเทือนถึงสิทธิอันชอบด้วยกฎหมายของเจ้าของลิขสิทธิ์เกินสมควร มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ภายใต้บังคับบทบัญญัติในวรรคหนึ่ง การกระทำอย่างหนึ่งอย่างใดแก่งานอันมีลิขสิทธิ์ตามวรรคหนึ่ง มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้กระทำดังต่อไปนี้

(๑) วิจัยหรือศึกษางานนั้น อันมิใช่การกระทำเพื่อหากำไร

(๒) ใช้เพื่อประโยชน์ของตนเอง หรือเพื่อประโยชน์ของตนเองและบุคคลอื่นในครอบครัวหรือญาติสนิท

(๓) ติชม วิจารณ์ หรือแนะนำผลงานโดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น

(๔) เสนอรายงานข่าวทางสื่อสารมวลชนโดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น

(๕) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ เพื่อประโยชน์ในการพิจารณาของศาลหรือเจ้าพนักงานซึ่งมีอำนาจตามกฎหมาย หรือในการรายงานผลการพิจารณาดังกล่าว

(๖) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ โดยผู้สอนเพื่อประโยชน์ ในการสอนของตน อันมิใช่การกระทำเพื่อหากำไร

(๗) ทำซ้ำ ดัดแปลงบางส่วนของงาน หรือตัดทอน หรือทำบทสรุปโดยผู้สอน หรือสถาบันศึกษาเพื่อแจกจ่ายหรือจำหน่ายแก่ผู้เรียนในชั้นเรียนหรือในสถาบันศึกษา ทั้งนี้ ต้องไม่เป็นการกระทำเพื่อหากำไร

(๘) นำงานนั้นมาใช้เป็นส่วนหนึ่งในการถามและตอบในการสอบ

มาตรา ๓๓ การกล่าว คัด ลอก เลียน หรืออ้างอิงงานบางตอนตามสมควรจากงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ โดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้ปฏิบัติตามมาตรา ๓๒ วรรคหนึ่ง

2.4 การป้อนข้อมูลเข้า

การป้อนข้อมูลเข้าสามารถทำได้สามวิธีดังที่กล่าวมาแล้ว คือ พิมพ์เข้าไป หรือใช้เอกสารอิเล็กทรอนิกส์ที่มีอยู่ในคอมพิวเตอร์อยู่แล้ว หรือสแกนเข้าไป ในวิธีแรกนั้น นักพิมพ์ดีดมืออาชีพที่พิมพ์สัมผัส 60-80 คำต่อนาที สามารถพิมพ์ภาษาอังกฤษได้ 10,000 คำในหนึ่งวัน (8 ชม.) ส่วนการใช้สแกนเนอร์นั้นก็สะดวกและมีคุณภาพมากขึ้น มีโปรแกรมรู้จำตัวอักษรหรือ OCR ภาษาอังกฤษที่ทำงานได้อย่างมีประสิทธิภาพสูง สามารถอ่านข้อมูลที่เป็นตัวพิมพ์ได้ถูกต้องดี ส่วนข้อมูลที่เป็นลายมือเขียนคงจะต้องนำเข้าโดยการพิมพ์เข้าไปใหม่ เพราะยังไม่มีโปรแกรมที่สามารถอ่านลายมือได้ถูกต้องอย่างมีประสิทธิภาพ

ส่วนข้อมูลภาษาพูดนั้นป้อนเข้าได้ยากกว่า เพราะต้องมีขั้นตอนเพิ่มเติม คือการถอดเสียงจากในเทปหรือวีดีโอลงในคอมพิวเตอร์ และการถอดเสียงเหล่านั้นออกมาเป็นคำถอดเสียง (transcription) ซึ่งเป็นที่รู้กันว่าเป็นเรื่องที่ใช้เวลามากเป็นอย่างยิ่ง อาจต้องใช้เวลาถึง 10-25 ชั่วโมงในการถอดเทป 1 ชั่วโมง ขึ้นอยู่กับว่าต้องการคุณภาพของการถอดเสียงในระดับไหน ส่วนคุณภาพของการอัดเสียงนั้นก็ขึ้นอยู่กับจุดประสงค์ว่าต้องการศึกษาในระดับไหน เพราะการศึกษา

ในบางเรื่องก็จำเป็นต้องมีการอัดเสียงเป็นอย่างดี เช่น การศึกษาเรื่องสัทสัมพันธ์ (prosody)

2.5 การกำกับข้อมูล

หากต้องการให้มีการใส่ข้อมูลอื่นๆที่นอกเหนือจากตัวเอกสารนั้นก็จำเป็นต้องมีการกำกับข้อมูลหรือใส่แท็ก แต่เดิมนั้น ต่างคนต่างก็กำหนดวิธีในการกำกับข้อมูลใช้เอง ต่อมามีการใช้ SGML (ISO 8879: 1986) เป็นมาตรฐานการกำกับ (encode) ข้อมูลอิเล็กทรอนิกส์ แต่ในปัจจุบัน XML ได้รับการยอมรับให้เป็นมาตรฐานแทน เอกสารที่กำกับข้อมูลตามมาตรฐานของ XML เป็นเอกสารที่สามารถนำไปใช้กับเครื่องคอมพิวเตอร์แบบใดก็ได้ ช่วยให้การแลกเปลี่ยนข้อมูลทางคอมพิวเตอร์เป็นไปได้โดยไม่มีข้อจำกัดด้านความแตกต่างของฮาร์ดแวร์หรือซอฟแวร์ นอกจากนี้ การกำหนดแท็กต่างๆขึ้นใช้ตามมาตรฐาน XML ควรเป็นไปตามแนวทางที่ TEI (Text Encoding Initiative) ได้วางเอาไว้ ซึ่ง TEI เป็นกลุ่มบุคคลที่มารวมตัวกันเป็นคณะกรรมการเพื่อกำหนดลักษณะทางโครงสร้างและแท็กต่างๆสำหรับใช้กำกับเอกสารประเภทต่างๆ การกำกับข้อมูลในคลังข้อมูลภาษาเป็นค่าใช้จ่ายที่เพิ่มขึ้นมาซึ่งผู้จัดทำจะต้องพิจารณาว่ามีความจำเป็นที่จะต้องใส่แท็กต่างๆโดยละเอียดมากน้อยเพียงใด และเป็นการลงทุนที่คุ้มกับประโยชน์ที่ได้หรือไม่

3. ตัวอย่างการจัดสร้างคลังข้อมูลภาษา

เพื่อที่จะให้เห็นภาพที่ชัดเจนของการจัดสร้างคลังข้อมูลภาษา ในที่นี้จะยกตัวอย่างของการออกแบบและจัดสร้างคลังข้อมูลสองตัวอย่างเพื่อแสดงให้เห็นว่ามีขั้นตอนการดำเนินการอย่างไร คลังข้อมูลภาษาที่ยกมาเป็นตัวอย่างในที่นี้

คือ คลังข้อมูล CSAE (Corpus of Spoken American English) (Chafe et al. 1991) และคลังข้อมูล BNC

3.1 คลังข้อมูล CSAE

ผู้สร้างคลังข้อมูลนี้ตั้งเป้าหมายว่าต้องการสร้างคลังข้อมูลภาษาพูดขนาด 200,000 คำของภาษาอังกฤษแบบอเมริกันซึ่งจะมีความยาวประมาณ 15 ชั่วโมง เพื่อเก็บไว้ในคอมพิวเตอร์ และให้ผู้ใช้สามารถค้นจากคำถอดเสียง (transcription) และฟังข้อมูลเสียงที่บันทึกไว้ได้พร้อมๆกัน

3.1.1 แรงจูงใจ

เกิดจากการที่ได้มีการสร้างคลังข้อมูลภาษาเขียนของภาษาอังกฤษแบบอังกฤษและคลังข้อมูลภาษาเขียนของภาษาอังกฤษแบบอเมริกันที่คล้ายคลึงกัน คือ คลังข้อมูล LOB และ Brown ตามลำดับ แต่ในส่วนของคลังข้อมูลภาษาพูดมีเพียงการสร้างคลังข้อมูลภาษาอังกฤษแบบอังกฤษ คือ London-Lund Corpus แต่ยังไม่มีผู้ใดสร้างคลังข้อมูลภาษาพูดของภาษาอังกฤษแบบอเมริกัน(ในขณะนั้น) ที่เก็บข้อมูลในลักษณะเดียวกันกับ London-Lund Corpus ดังนั้น จึงน่าจะมีการจัดสร้างคลังข้อมูล Corpus of Spoken American English (CSAE) เพื่อจะได้มีคู่เทียบที่สมบูรณ์ระหว่างภาษาอังกฤษแบบอังกฤษและภาษาอังกฤษแบบอเมริกัน

3.1.2 ประโยชน์ที่คาดไว้

- เป็นข้อมูลสำหรับผู้ที่จะศึกษาบทสนทนา ภาษาพูด
- เป็นประโยชน์สำหรับผู้ที่สอนภาษาอังกฤษเป็นภาษาที่สอง (ESL)
- เป็นหลักฐานภาษาพูดของภาษาในปลายศตวรรษที่ 20

- ข้อมูลจริงจะเปิดเผยให้เห็นถึงแง่มุมและประเด็นต่างๆที่ทฤษฎีทางภาษาศาสตร์ยังไม่ได้ศึกษา

3.1.3 สิ่งที่คำนึงถึงในการสร้างคลังข้อมูล CSAE

1. สามารถเปรียบเทียบได้ (Comparability)

ถ้าสามารถที่จะสร้างคลังข้อมูลภาษาโดยใช้ต้นแบบจากคลังข้อมูลภาษาอื่นที่มีมาก่อน ก็จะง่ายและสะดวกต่อการเปรียบเทียบผลที่ได้จากคลังข้อมูลภาษาทั้งสอง แต่ในขณะเดียวกัน แต่ละคลังข้อมูลภาษาก็ถูกสร้างเพื่อวัตถุประสงค์ที่ต่างกัน เทคโนโลยีในเวลาที่สร้างก็แตกต่างกัน จึงอาจจะไม่ง่ายนักที่จะสร้างคลังข้อมูลภาษาโดยใช้โครงสร้างที่เหมือนกันทีเดียว

ในกรณีของคลังข้อมูล CSAE นี้คลังข้อมูลภาษาที่มีการสร้างมาก่อนที่สามารถเทียบกันได้คือคลังข้อมูล London-Lund อย่างไรก็ตาม คลังข้อมูล CSAE ไม่ได้เก็บข้อมูลเหมือนคลังข้อมูล London-Lund ในหลายประเด็น เช่น เรื่องขนาด คลังข้อมูล CSAE เก็บคำ 200,000 คำในขณะที่คลังข้อมูล London-Lund เก็บ 500,000 คำ ทั้งนี้เนื่องจากผู้สร้างคลังข้อมูล CSAE ต้องการบันทึกข้อมูลไว้ให้ได้ในซีดีรอมหนึ่งแผ่นแทนที่จะต้องเก็บไว้ในซีดีรอมหลายแผ่น อีกทั้งเรื่องเวลาและงบประมาณก็เป็นอุปสรรคหนึ่งที่สำคัญ การเก็บข้อมูลเพียงแค่ 200,000 คำนี้ก็คาดว่าจะใช้เวลาถึง 3 ปี ส่วนในเรื่องรูปแบบ คลังข้อมูล CSAE ไม่ได้จำกัดว่าขนาดของตัวอย่างต้องเท่ากันเสมอเหมือนอย่างในคลังข้อมูล London-Lund (อาจเป็นไปได้ว่า ในสมัยที่มีการจัดสร้างคลังข้อมูล London-Lund ต้องเก็บข้อมูลลงไว้ในเทปคอมพิวเตอร์จึงจำเป็นต้องกำหนดให้แต่ละระเบียนมีขนาดแน่นอน (fixed length record)) ข้อมูลที่เป็นการสนทนาในคลังข้อมูล CSAE จึงมีอยู่ครบ ซึ่งจะเป็นประโยชน์ต่อการศึกษาทางภาษาศาสตร์มากกว่า ส่วนในเรื่องของการถอดเสียงของคลังข้อมูล CSAE ใช้วิธีการถอดเสียง

ที่ง่ายกว่าคลังข้อมูล London-Lund ทั้งนี้เพื่อให้ผู้ใช้สามารถใช้ได้โดยสะดวก แต่ก็พยายามถอดข้อมูลเสียงที่สำคัญของภาษาพูดไว้ให้ครบ

อย่างไรก็ตาม ผู้สร้างคลังข้อมูล CSAE ได้ชี้ให้เห็นว่าคลังข้อมูล CSAE สามารถเปรียบเทียบกับคลังข้อมูล London-Lund ได้ ทั้งนี้เพราะ ในคลังข้อมูล London-Lund นั้นมีข้อมูลการสนทนาเพียง 250,000 คำ ที่เหลือเป็นข้อมูลภาษาพูดด้านอื่น อีกทั้งนักวิจัยส่วนมากที่ใช้คลังข้อมูล London-Lund ก็มักจะใช้ข้อมูลส่วนที่เป็นการสนทนาเป็นส่วนมาก ดังนั้น คลังข้อมูล CSAE จึงถูกออกแบบให้เก็บเฉพาะการสนทนา 200,000 คำ ซึ่งก็ใกล้เคียงกับข้อมูลการสนทนา 250,000 คำของคลังข้อมูล London-Lund

2. การสุ่มตัวอย่าง

ผู้สร้างตั้งเป้าหมายว่าคลังข้อมูลภาษานี้จะเป็นภาษาที่เป็นภาษามาตรฐานของภาษาอังกฤษแบบอเมริกัน และจะให้มีความหลากหลายทางด้านอายุ เพศ ภูมิลำเนา อาชีพ การศึกษา สถานการณ์ แต่คลังข้อมูลภาษานี้ไม่ได้ตั้งเป้าเพื่อใช้สำหรับการศึกษาด้านภาษาศาสตร์สังคม (sociolinguistics) ซึ่งเป็นงานที่ต้องการการเก็บข้อมูลของผู้พูด (speaker) ในกลุ่มสังคมหนึ่ง (one community) เป็นจำนวนมาก ความเป็นภาษามาตรฐานไม่ได้จำกัดอยู่ที่จะต้องเป็นภาษาพูดหรือภาษาเขียนที่ปรากฏตามวารสาร แต่ในที่นี้พิจารณาจากลักษณะของภาษาโดยดูว่าเป็นไปตามกฎทางไวยากรณ์และโครงสร้างที่เหมือนๆกัน ซึ่งจะเลือกข้อมูลที่ใกล้เคียงกับภาษาเขียนของภาษาอังกฤษแบบอเมริกันที่พบตามวารสาร และจะไม่ตั้งข้อจำกัดในเรื่องของสำเนียงหรือสไตล์การใช้ภาษา

ข้อมูลประกอบด้วยการสนทนาแบบเผชิญหน้า (face-to-face) ความยาวครึ่งชั่วโมงจำนวน 30 เหตุการณ์ และการสนทนาสั้นๆจำนวนหนึ่งแต่รวมกัน

ไม่เกินร้อยละ 5 ของคลังข้อมูลภาษา ซึ่งจะเห็นว่าข้อมูลแค่นี้ไม่เพียงพอที่จะสะท้อนความแตกต่างระหว่างภาษาถิ่นต่างๆหรือระหว่างกลุ่มชาติพันธ์ (ethnic group) ต่างๆ งานทางภาษาศาสตร์สังคมโดยปกติจะเก็บข้อมูลมากกว่านี้ เช่น เก็บข้อมูลเป็นร้อยชั่วโมง แต่ไม่ได้จำเป็นที่จะต้องถอดเสียงคำทุกคำในข้อมูลที่เก็บมา เพียงแต่เลือกถอดเสียงเฉพาะส่วนที่มีข้อมูลที่ต้องการเท่านั้น แต่ในคลังข้อมูล CSAE นี้ คำพูดทุกๆคำจะต้องถูกถอดเสียงโดยละเอียดเพื่อจะเป็นประโยชน์ต่อการศึกษาในแง่มุมใดก็ได้

อีกเหตุผลหนึ่งที่เลือกภาษามาตรฐานของภาษาอังกฤษแบบอเมริกัน ก็เพื่อจะได้เปรียบเทียบคลังข้อมูล CSAE กับคลังข้อมูลภาษาอื่นซึ่งเก็บภาษามาตรฐานที่เป็นภาษาเขียนของภาษาอังกฤษแบบอเมริกันหรือเพื่อเปรียบเทียบกับคลังข้อมูลที่เป็นภาษามาตรฐานของภาษาพูดภาษาอังกฤษแบบอังกฤษได้ นอกจากนี้ CSAE ยังจะมีประโยชน์เพื่อใช้ในการสอนภาษาอังกฤษเป็นภาษาที่สองได้อีกด้วย

3.1.4 ขั้นตอนของการสร้าง CSAE

- บันทึกบทสนทนาจากสถานการณ์จริงโดยใช้เครื่องบันทึกเสียงแบบพกพาและไมโครโฟนคุณภาพสูงสำหรับงานภาคสนาม
- จดบันทึกข้อมูลเกี่ยวกับผู้สนทนา สถานที่ และสภาพแวดล้อมทั่วไป
- ทำสำรองเทปที่ได้เพื่อป้องกันการสูญหาย
- คัดเลือกข้อมูลจากเทปเพื่อใส่ในคลังข้อมูล CSAE ให้ได้ประมาณ 15 ชั่วโมง
- ถอดเสียงในเทปที่ได้เลือกไว้
- มีการตรวจสอบข้อมูลที่ถอดเสียงออกมาจนกระทั่งพอใจ
- เสียงในเทปจะถูกนำถ่ายเข้าในคอมพิวเตอร์ในรูปของซีดีรอม
- จัดทำโปรแกรมดัชนีสำหรับเชื่อมโยงคำถอดเสียงกับเสียงที่บันทึกอยู่ในซีดีรอม

- เผยแพร่คลังข้อมูลภาษาในรูปของซีดีรอมและหนังสือ ทั้งนี้เพราะคนบางคนอาจไม่มีเครื่องอ่านซีดี หรือคนบางคนก็ชอบที่จะดูข้อมูลผ่านทางหนังสือมากกว่าบนหน้าจอคอมพิวเตอร์ คลังข้อมูลในรูปของซีดีรอมกำหนดขายในราคา 80 ดอลลาร์สหรัฐ โดยที่หนังสือจะมีเฉพาะคำถอดเสียงแบบหยาบๆ ในขณะที่ซีดีรอมจะมีคำถอดเสียงทั้งแบบหยาบและแบบละเอียด คำถอดเสียงแบบหยาบจะง่ายต่อการทำความเข้าใจของคนส่วนมาก แต่นักวิจัยบางคนอาจจำเป็นต้องการใช้คำถอดเสียงแบบละเอียดมากกว่า

3.2 คลังข้อมูล BNC

3.2.1 ลักษณะทั่วไป

คลังข้อมูล BNC ถูกสร้างขึ้นมาเพื่อเป็นคลังข้อมูลภาษาอังกฤษแบบอังกฤษในปัจจุบัน โดยต้องการให้มีขนาดใหญ่มากกว่า 100 ล้านคำ หรือกินเนื้อที่ประมาณ 1.5 กิกะไบท์ เพื่อให้ใช้ประโยชน์ในงานต่างๆ เช่น งานทางด้านการประมวลผลภาษาธรรมชาติ งานด้านภาษาศาสตร์ หรือใช้เป็นตัวอย่างจริงของการใช้ภาษาปัจจุบันสำหรับการทำพจนานุกรม การสอนภาษาอังกฤษ

BNC จัดทำขึ้นภายใต้การนำของสำนักพิมพ์ Oxford University Press โดยอาศัยความร่วมมือจากหลายๆหน่วยงาน ได้แก่ สำนักพิมพ์ Addison-Wesley สำนักพิมพ์ Longman และสำนักพิมพ์ Larousse Kingfisher Chambers ศูนย์วิจัยที่ Oxford University Computing Services, Centre for Computer Research on the English Language ของมหาวิทยาลัย Lancaster และ Research and Innovation Centre ของ British Library โครงการจัดทำ BNC เริ่มในปี 1991 และเสร็จสิ้นในเดือนกุมภาพันธ์ปี 1995 ราคาจำหน่ายของ BNC เมื่อเริ่มต้นเผยแพร่อยู่ที่ 220 ปอนด์ และมีค่าสิทธิการใช้ (license) 10 ปอนด์

สำหรับ 5 ปี โดยผู้ซื้อจะได้ทั้งคลังข้อมูลภาษา หนังสือคู่มือและโปรแกรม SARA สำหรับใช้สืบค้นข้อมูลใน BNC คลังข้อมูลที่จำหน่ายจะอยู่ในรูปของซีดีรอม 2 แผ่น หรืออาจใช้ผ่านทางอินเทอร์เน็ตโดยเสียค่าสมาชิกรายปี ปีละ 60-500 ปอนด์ตามลักษณะเงื่อนไขต่างๆ ในช่วงเริ่มแรกนี้ ยังไม่มีการจำหน่ายคลังข้อมูลนี้แก่ประเทศนอกทวีปยุโรปเนื่องจากข้อจำกัดด้านลิขสิทธิ์ แต่ปัจจุบัน ได้จำหน่ายแก่ผู้สนใจทั่วโลกในชื่อของคลังข้อมูล BNC Word Edition โดยปรับราคาจำหน่ายอยู่ที่ 50 ปอนด์สำหรับผู้ใช้คนเดียว และ 250 ปอนด์สำหรับใช้ในเครือข่ายเน็ตเวิร์ค

BNC เป็นคลังข้อมูลที่ถูกออกแบบให้ครอบคลุมภาษาในลักษณะต่างๆ ให้มากที่สุด ประกอบด้วยตัวบท 4,124 ชิ้นจากเอกสารประเภทต่างๆ ซึ่งแบ่งเป็นตัวบทภาษาเขียนร้อยละ 90 ตัวบทภาษาพูดร้อยละ 10 ข้อมูลใน BNC เป็นข้อมูลภาษาที่มีการแบ่งประโยคและแบ่งคำ แต่ละคำมีการกำกับหมวดคำซึ่งได้จากการใช้โปรแกรม CLAW ของมหาวิทยาลัย Lancaster ในการกำกับหมวดคำ ข้อมูลถูกกำกับตามมาตรฐานที่กำหนดโดย TEI ซึ่งเป็นไปตามมาตรฐาน ISO standard 8879 หรือคือมาตรฐาน SGML (ในปัจจุบัน ได้ปรับการกำกับข้อมูลโดยใช้ XML แทน SGML) ข้อมูลต่างๆเกี่ยวกับเอกสารที่นำมาใช้ เช่น ข้อมูลบริบทที่ใช้ (contextual) และข้อมูลบรรณานุกรม (bibliography) จะถูกเก็บไว้ในส่วนที่เป็น TEI header

ในกรณีที่ผู้ใช้ไม่ต้องการใช้คลังข้อมูลทั้งหมดขนาดหนึ่งร้อยล้านคำ ก็อาจเลือกใช้เฉพาะส่วนที่เรียกว่าส่วนแกนหลักของคลังข้อมูลซึ่งมีขนาด 2 ล้านคำ โดยมีสัดส่วนของตัวบทประเภทต่างๆเหมือนกับตัวคลังข้อมูลใหญ่ แต่ในคลังข้อมูลหลักนี้สัดส่วนระหว่างข้อมูลภาษาเขียนและภาษาพูดจะมีเท่าๆกัน คือ ร้อยละ 50

3.2.2 การคัดเลือกข้อมูล

3.2.2.1 การคัดเลือกข้อมูลเอกสารภาษาเขียน

การคัดเลือกข้อมูลภาษาเขียนทำโดยใช้เกณฑ์หลัก 3 อย่าง คือ เนื้อหา (domain) เวลา และ สื่อที่ใช้

เนื้อหา เป็นเกณฑ์ที่ใช้เพื่อแยกประเภทของงานเขียน โดยกำหนดให้ร้อยละ 75 ของงานเขียนที่ต้องการเป็นข้อเขียนด้านข้อมูลข่าวสาร (informative) ซึ่งเลือกจากสาขาต่างๆ เช่น applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs และอีกร้อยละ 25 ของงานเขียนเป็นเอกสารด้านจินตนาการ (imaginative) ซึ่งคืองานด้านวรรณกรรมและงานสร้างสรรค์ต่างๆ

สื่อที่ใช้ เป็นเกณฑ์ที่ใช้เพื่อแยกประเภทของเอกสารที่ปรากฏในสื่อ โดยกำหนดให้ร้อยละ 60 ของงานเขียนเป็นหนังสือ ร้อยละ 25 เป็นวารสารและหนังสือพิมพ์ อีกร้อยละ 5-10 มาจากสิ่งพิมพ์อื่นๆ เช่น แผ่นพับ (brochures) ใบปลิวโฆษณา (advertising leaflets) อีกร้อยละ 5-10 มาจากงานเขียนที่ไม่มีการตีพิมพ์ เช่น จดหมายส่วนตัว บันทึกประจำวัน (diaries) ความเรียง (essays) และบันทึกช่วยจำ (memo) และอีกน้อยกว่า 5% เป็นเอกสารที่เขียนขึ้นสำหรับใช้ในการพูด เช่น คำปราศรัย บทละคร บทสำหรับอ่านออกอากาศ (broadcast scripts)

เวลา เป็นเกณฑ์ที่ใช้เพื่อแยกประเภทของเอกสารตามเวลาที่ตีพิมพ์ โดยกำหนดว่า จะใช้เอกสารตั้งแต่ปี 1975 ไปต้นไป ยกเว้นเอกสารด้านจินตนาการ เช่น นิยายซึ่งยอมใช้เอกสารเก่าถึงปี 1964 ได้เพราะนิยายบางเล่มยังเป็นที่นิยมอ่านกันอยู่

นอกจากเกณฑ์หลักทั้งสามแล้ว ยังมีเกณฑ์ย่อยอื่นๆอีก แต่ไม่ได้กำหนดกฎเกณฑ์ไว้ตายตัวว่าต้องเป็นเช่นไร เพียงแค่พยายามให้มีความหลากหลายเท่าที่จะเป็นไปได้ในเกณฑ์นั้น ๆ เช่น ขนาดของเอกสารและขอบเขตข้อมูล (จุดเริ่มต้นและจุดสิ้นสุด) หัวข้อเรื่องของเอกสาร ชื่อ อายุ เพศ และภูมิลำเนาของผู้แต่ง อายุและเพศของกลุ่มเป้าหมาย ระดับความยากง่ายของเอกสาร เป็นต้น

3.2.2.2 การคัดเลือกข้อมูลเอกสารภาษาพูด

เอกสารภาษาเขียนแยกออกเป็นสองส่วนเท่าๆกัน คือ ส่วนที่เป็นภาษาพูดแยกตามภูมิภาค (demographic) ซึ่งเป็นส่วนที่เป็นการถอดเสียงของบทสนทนาต่างๆของอาสาสมัครจากภูมิลำเนาต่างๆ และส่วนที่เป็นภาษาพูดแยกตามบริบท (context-governed) ซึ่งเป็นส่วนที่มาจากการถอดเสียงของเทปบันทึกการประชุมต่างๆ ที่ได้จากสถานการณ์แบบต่างๆ

ข้อมูลที่เป็นภาษาพูดแยกตามภูมิภาค ได้มาจากอาสาสมัคร 124 คน จากกลุ่มสังคมต่างๆ ทั้งชาย และหญิงอายุต่างๆ อาสาสมัครมาจากภูมิลำเนา 38 แห่งทั่วสหราชอาณาจักร การคัดเลือกพยายามให้มีเพศชายหญิงเท่ากัน มีจำนวนคนในแต่ละกลุ่มอายุเท่าๆกัน และในแต่ละกลุ่มสังคมเท่าๆกัน โดยที่อาสาสมัครจะอัดเทปการสนทนาของตัวเองที่เกิดขึ้นในชีวิตประจำวันเป็นเวลา 2-3 วัน พร้อมทั้งจดบันทึกของบทสนทนาแต่ละบทไว้ว่าเกิดขึ้นที่ไหนกับใครและเมื่อไร

ส่วนข้อมูลที่เป็นภาษาพูดแยกตามบริบท ได้ตั้งเป้าหมายว่าจะเก็บคำพูดต่อเนื่องที่บันทึกไว้ในสี่สาขาในปริมาณเท่าๆกัน ดังต่อไปนี้

1. ด้านการศึกษาหรือการให้ข่าวสารข้อมูล เช่น การบรรยายในห้องเรียน การรายงานข่าว การอภิปรายในชั้นเรียน การอบรม

2. ด้านธุรกิจ เช่น การสาธิตประกอบการขาย การประชุมการค้า การให้คำปรึกษา การสัมภาษณ์

3. ด้านองค์กรและสาธารณะ เช่น การกล่าวสวด การกล่าวปราศรัย การประชุมกรรมการ การประชุมสภา

4. ด้านบันเทิง เช่น การบรรยายกีฬา การประชุมในคลับ การสนทนาผ่านรายการวิทยุ

3.2.3 การสร้างคลังข้อมูล BNC

มีขั้นตอนใหญ่ๆ อยู่ 5 ขั้นตอน คือ

1. การขออนุญาตรวบรวมเอกสาร ขั้นตอนนี้อยู่ในความรับผิดชอบของสำนักพิมพ์ Oxford University Press โดยใช้วิธีร่างแบบฟอร์มข้อตกลงมาตรฐาน โดยปรึกษากับสมาคมนักเขียน (The Society of Authors) และสมาคมผู้ผลิตสิ่งพิมพ์ (the Publishers Association) และกลุ่มอื่นๆที่เกี่ยวข้อง ที่ต้องทำเช่นนี้เพราะการทำข้อตกลงเป็นรายๆสำหรับเอกสารแต่ละชิ้นจะทำให้ต้องเสียค่าใช้จ่ายสูงมาก ประเด็นสำคัญที่อยู่ในข้อตกลง คือ

เจ้าของลิขสิทธิ์ยินยอมให้รวมเอกสารนั้นโดยไม่คิดค่าใช้จ่าย

โครงการที่จัดทำจะไม่มุ่งแสวงหากำไรในเชิงพาณิชย์

โครงการนี้สร้างข้อตกลงว่า สำเนาของคลังข้อมูลจะจำหน่ายภายใต้ข้อตกลงมาตรฐานนี้ด้วย

โครงการนี้จะไม่บรรจุเอกสารฉบับเต็มในเรื่องนั้นๆ ลงในคลังข้อมูล

2. การรวบรวมเอกสาร กระทำโดยความร่วมมือของหลายหน่วยงาน สำนักพิมพ์ Oxford University Press เป็นผู้รับผิดชอบหาตัวบทภาษาเขียน

Longman Group รับผิดชอบด้านตัวบทภาษาพูด Chambers รับผิดชอบรวบรวมเอกสารที่ไม่มีการตีพิมพ์ การนำข้อมูลลงในคอมพิวเตอร์ ทำโดยใช้ 3 วิธี คือ การสแกน การพิมพ์โดยตรง และการใช้เอกสารอิเล็กทรอนิกส์ที่มีอยู่แล้ว

- การสแกน เอกสารที่สแกนเข้าเลือกเฉพาะที่มีภาพตัวอักษรที่คมชัด หลังจากสแกนแล้วก็ต้องให้คนตรวจสอบแก้ไขอีกที และกำกับข้อมูลเพิ่มเติมโดยการใส่แท็กต่างๆ
- การพิมพ์เข้า ใช้สำหรับเอกสารที่เป็นลายมือเขียน หรือเอกสารที่ตัวพิมพ์เล็กเกินกว่าที่เครื่องสแกนเนอร์จะอ่านเอกสารได้ชัดเจน ซึ่งการพิมพ์เอกสารใหม่เลยจะเร็วกว่าการให้คนมาแก้ไขเอกสารที่สแกนแล้วมีข้อผิดพลาดมาก
- การใช้เอกสารอิเล็กทรอนิกส์ที่มีอยู่แล้ว ซึ่งน่าจะมีอยู่เป็นจำนวนมาก เพียงแค่นำมาปรับให้เข้ากับรูปแบบมาตรฐานที่กำหนดก็น่าจะใช้การได้ดี แต่ในความเป็นจริง เอกสารอิเล็กทรอนิกส์ที่ตรงตามที่ออกแบบไว้กลับไม่ได้มีมากดังที่คาด และเมื่อได้มาแล้วก็ต้องใช้โปรแกรมช่วยในการจัดรูปแบบใหม่

3. การกำกับข้อมูลภาษา ขั้นตอนนี้รับผิดชอบโดย OUCS (Oxford University Computing Service) ขั้นตอนนี้เป็นการแปลงเอกสารที่ได้มาให้เป็นไปตามมาตรฐานของการเข้ารหัสข้อมูลตามมาตรฐานที่กำหนดขึ้น (เอกสารใน BNC ถูกกำกับข้อมูลต่างๆ เช่น ขอบเขตคำ หมวดคำ ขอบเขตประโยค ย่อหน้า ตอน หัวเรื่อง ในภาษาพูด ก็จะมีข้อมูลอื่นๆด้วย เช่น ผลัดของการสนทนา การหยุด สัทลักษณ์อื่นๆที่ไม่เกี่ยวกับภาษา (para-linguistic features) เช่น เสียงหัวเราะ และยังมีข้อมูลที่เกี่ยวกับตัวบท (meta-textual information) นั้นๆ เช่น แหล่งที่มา วิธีการกำกับข้อมูล ฯลฯ มาตรฐานที่กำหนดขึ้นสำหรับกำกับข้อมูลในคลังข้อมูล BNC เรียกว่า Corpus Document Interchange Format (CDIF) CDIF เป็นการกำกับข้อมูลที่ประยุกต์ใช้ตามมาตรฐาน SGML (ISO 8879) จึงทำ

ให้เราสามารถใช้โปรแกรมใดใดที่เข้าใจรูปแบบของ SGML กับคลังข้อมูล BNC นี้ได้ การกำหนด CDIF ก็เป็นไปตามแนวทางที่ TEI วางไว้ ดังนั้น จึงมั่นใจได้ว่าคลังข้อมูล BNC จะสามารถนำไปใช้กับคอมพิวเตอร์แบบใดหรือซอฟท์แวร์ใดก็ได้

4. การกำกับข้อมูลทางภาษาศาสตร์ ขั้นตอนนี้อยู่ในความรับผิดชอบของ UCREL (the Unit for Computer Research on the English Language) ขั้นตอนนี้เป็นการใส่หมวดคำให้กับคำแต่ละคำในตัวบท โดยอาศัยโปรแกรมกำกับหมวดคำ CLAW ของมหาวิทยาลัย Lancaster แท็กพื้นฐานที่ใช้ในการกำกับหมวดคำของ BNC มี 61 หมวดคำ แต่ในส่วนที่เป็นแกนหลักของ BNC ขนาด 2 ล้านคำจะใช้แท็กที่ละเอียดกว่าซึ่งมีมากกว่า 160 หมวดคำ เนื่องจากคลังข้อมูลนี้มีขนาดใหญ่มากจึงไม่สามารถตรวจสอบว่าหมวดคำที่กำกับให้กับคำทั้งหมดนั้นถูกต้อง ต้องยอมให้มีการกำกับหมวดคำที่ผิดพลาดแฝงอยู่ในคลังข้อมูลบ้าง ซึ่งกระบวนการกำกับหมวดคำโดยอัตโนมัตินี้มีความผิดพลาดประมาณร้อยละ 1.7 โดยที่มีคำประมาณร้อยละ 4.7 ที่ไม่สามารถกำกับหมวดคำที่แน่นอนได้ (มีหมวดคำมากกว่าหนึ่ง)

5. การจัดเก็บตัวบท เป็นขั้นตอนสุดท้ายในความรับผิดชอบของ OUCS (Oxford University Computing Service) ขั้นตอนนี้เป็นการใส่ข้อมูลเกี่ยวกับตัวบทนั้นๆไว้ในส่วนที่เป็นส่วนหัว (header) และตรวจสอบโครงสร้างของตัวบทเป็นครั้งสุดท้ายว่าเป็นไปตามมาตรฐาน SGML หรือไม่ ซึ่งอาจทำให้ต้องมีการแก้ไขด้วยมือในตอนหลังบ้าง แต่เป็นปริมาณที่ไม่มากนัก ประมาณร้อยละ 5-10 ของข้อมูลทั้งหมด

บทที่ 9 การนำข้อมูลภาษาเข้าคอมพิวเตอร์

การสร้างคลังข้อมูลเป็นการนำข้อมูลภาษาลงจัดเก็บในคอมพิวเตอร์ ซึ่งสามารถทำได้หลายวิธี วิธีที่ง่ายแต่สิ้นเปลืองที่สุดคือการพิมพ์ข้อมูลเข้าไปใหม่ แต่หากเป็นข้อมูลที่มีผู้นำเข้าไว้ในคอมพิวเตอร์แล้ว ก็จะสามารถประหยัดค่าใช้จ่ายได้ด้วยการคัดลอกหรือแปลงไฟล์ข้อมูลที่มีอยู่แล้วเข้าไปเป็นส่วนหนึ่งของคลังข้อมูล บทนี้จะเริ่มด้วยการกล่าวตั้งแต่เรื่องพื้นฐานของการเก็บข้อมูลในทางคอมพิวเตอร์ การกำหนดรหัสอักขระสำหรับภาษา การเก็บข้อมูลไฟล์ในรูปแบบต่างๆ และการแปลงข้อมูลไฟล์ต่างๆ มาเป็นไฟล์ข้อความล้วน ในส่วนหลังจะกล่าวถึงกรณีที่ข้อมูลที่ต้องการมีเฉพาะที่พิมพ์บนกระดาษ ซึ่งเราก็สามารถนำข้อมูลภาษาเข้าโดยใช้เครื่องสแกนเน่อร์ร่วมกับโปรแกรมรู้จำตัวอักษรได้ หรือในกรณีที่ข้อมูลที่ต้องการมีการจัดเก็บไว้อย่างเป็นระบบในเว็บไซต์ต่างๆ เราก็อาจใช้โปรแกรมคัดลอกข้อมูลทั้งหมดมาเพื่อนำมาใช้ได้

1. การเก็บข้อมูลในคอมพิวเตอร์

พื้นฐานของเรื่องนี้คือการเข้าใจว่าข้อมูลเข้าไปเก็บในคอมพิวเตอร์ได้อย่างไร คอมพิวเตอร์เป็นอุปกรณ์อิเล็กทรอนิกส์ มีลักษณะเป็นดิจิตอล ดังนั้นพื้นฐานเล็กสุดคือการมีไฟหรือไม่มีไฟ หรือสถานะเปิดหรือปิด ซึ่งทางคอมพิวเตอร์เรียกว่า บิต ซึ่งมีได้สองสถานะคือเป็น 1 หรือ 0 การมีบิตแค่ศูนย์และหนึ่งก็ไม่น่าจะทำให้คอมพิวเตอร์เก็บข้อมูลอะไรได้ แต่หากเราใช้จำนวนบิตมากขึ้น เช่น ถ้านำสองบิตมาต่อกัน ความเป็นไปได้ของสถานะก็จะเพิ่มขึ้น ได้แก่ 00, 01, 10, 11 จำนวนที่สองบิตจึงใช้แทนสิ่งที่ต่างกันได้สี่อย่าง และหากเพิ่ม

จำนวนบิตมากขึ้นจำนวนของสิ่งที่จะแทนได้ก็มากขึ้นเป็นทวีคูณ เช่น ใช้สามบิตจะแทนข้อมูลได้แปดอย่าง ใช้สี่บิตแทนจะข้อมูลได้สิบหกอย่าง ดังนั้น หากเราใช้สามบิตแทนข้อมูลภาษาเพื่อเก็บลงในคอมพิวเตอร์ เราก็จะเก็บตัวอักษรที่ต่างกันได้ 8 ตัวอักษร ดังตารางสมมติข้างล่าง

รหัสข้อมูลที่เก็บ	ตัวอักษร
0 0 0	A
0 0 1	B
0 1 0	C
0 1 1	D
1 0 0	E
1 0 1	F
1 1 0	G
1 1 1	H

ตาราง 9.1 ตัวอย่างการแทนอักษรด้วยเลขสามบิต

การนำข้อมูลภาษาเข้าไปเก็บในคอมพิวเตอร์ก็คือการแทนที่ด้วยบิตในคอมพิวเตอร์นี้ ซึ่งความจริงก็ไม่แตกต่างอะไรกับการที่มนุษย์สร้างคำต่างๆ ขึ้นมากมายจากการผสมของหน่วยเสียงที่มีจำนวนจำกัด เพียงแต่คอมพิวเตอร์มีเพียงสองหน่วยคือ 0 กับ 1 ให้มาผสมกันเท่านั้น คำถามที่ต้องพิจารณาต่อ คือ ควรใช้จำนวนบิตเท่าใดมาผสมกันจึงจะแทนหรือเก็บข้อมูลภาษามนุษย์ได้เพียงพอ หากพิจารณาภาษาในลักษณะที่เป็นภาษาเขียนแบบที่ประกอบด้วยตัวอักษรต่างๆ แต่ละภาษาก็อาจมีตัวอักษรไม่เท่ากัน ภาษาอังกฤษมีตัวอักษร a-z 26 ตัวอักษรและมีทั้งที่เป็นตัวพิมพ์ใหญ่และตัวพิมพ์เล็ก จึงมี 52 ตัวอักษร และก็มีตัวเลข 0-9 และเครื่องหมายต่างๆ เช่น : ; , . “ + - = > < ? เป็นต้น ซึ่ง

รวมกันแล้วมากกว่า 64 ตัวอักษรแต่น้อยกว่า 128 ตัวอักษร ข้อมูลภาษาอังกฤษจึงสามารถแทนได้ด้วยรหัสจำนวน 7 บิต (2^7=128)

2. รหัสอักขระ

ในการกำหนดว่า ตัวอักขระหรือตัวอักษรใดแทนด้วยรหัสใดนั้นก็ต้องตกลงให้เป็นมาตรฐานร่วมกันด้วย รหัสที่ใช้แทนตัวอักษรภาษาอังกฤษที่เป็นที่ยอมรับใช้กันในปัจจุบัน คือ รหัส ASCII ย่อมากจาก American Standard Code for Information Interchange ซึ่งเผยแพร่ให้ใช้ครั้งแรกตั้งแต่ปี ค.ศ.1963 โดยหน่วยงานที่ปัจจุบันมีชื่อว่า ANSI (American National Standards Institute) และมีการแก้ไขครั้งใหญ่ในปีค.ศ.1967 และปรับแก้เป็นแบบที่ใช้ในปัจจุบันในปี ค.ศ.1986 ในจำนวนรหัส 128 ตัวนั้น 95 ตัวใช้แทนอักษรที่มองเห็นส่วนอีก 33 ตัวใช้แทนรหัสควบคุมในทางคอมพิวเตอร์ที่ไม่เกี่ยวกับตัวอักษรในภาษา รูปข้างล่างแสดงตารางรหัส ASCII นี้ (รหัส 0-32 ไม่แสดงไว้เพราะแทนสิ่งที่ไม่ใช่ตัวอักษรใดใด) และเพื่อความสะดวกสำหรับมนุษย์ แทนที่เราจะเขียนด้วยเลขบิต 0 1 เราก็จะเขียนเป็นเลขฐานสิบธรรมดา รหัส 0110011 จึงเขียนเป็น 51, 0110100 เขียนเป็น 52 เป็นต้น หรือบางครั้งนักคอมพิวเตอร์ก็นิยมเขียนเป็นเลขฐานสิบหกคือเขียนเป็น 33, 34 ตามลำดับ

Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(	56	8	72	H	88	X	104	h	120	x
41	)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[	107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	|
45	-	61	=	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	_
48	0	64	@	80	P	96	`	112	p		

ตาราง 9.2 ตารางรหัส ASCII นับตามเลขฐานสิบ

2.1 รหัสอักขระภาษาต่าง ๆ

ตารางรหัส ASCII แทนข้อมูลที่เป็นภาษาอังกฤษได้ แต่ในภาษาอื่นๆ นั้นก็จะมีตัวอักษรที่ต่างออกไป แม้แต่ภาษากลุ่มยุโรปที่ใช้ตัวอักษรโรมัน ก็จะมีการใช้อักขระอื่นนอกเหนือจากที่ใช้ในภาษาอังกฤษ เช่น ß (German), ñ (Spanish), å (Swedish) จึงได้มีการเพิ่มเติมรหัสสำหรับตัวอักษรอื่นๆ เรียกว่า extended ASCII การที่จะเพิ่มได้ก็คือเพิ่มจำนวนบิตที่ใช้แทนข้อมูลอีกหนึ่งบิต จาก 7 บิตเป็น 8 บิต เมื่อเป็น 8 บิตแล้ว (หรือในทางคอมพิวเตอร์เรียก 1 byte) จำนวนตัวอักษรที่แทนได้ทั้งหมดจึงเพิ่มจาก 128 ตัวอักษรเป็น 256 ตัวอักษร (2^8) ซึ่งเครื่องคอมพิวเตอร์ส่วนบุคคลเมื่อเริ่มใช้ 8 บิตสำหรับเก็บรหัสตัวอักษร ในระยะแรก บริษัทคอมพิวเตอร์รายใหญ่แต่ละแห่งก็มีการกำหนดรหัสที่แตกต่าง

กัน เช่น รหัส IBM-PC ของบริษัทไอบีเอ็ม, รหัส DEC ของบริษัท Digital Equipment Corporation, รหัส Mac Roman ของบริษัท Apple จนกระทั่งหน่วยงาน ISO (International Organization for Standardization) ได้กำหนดมาตรฐานรหัสอักขระส่วนเพิ่มเติมนี้สำหรับกลุ่มภาษาต่างๆ จนกระทั่งมาเป็นรหัสอักขระมาตรฐานที่รู้จักกันในชื่อ ISO-8859-x โดยที่ x แทนกลุ่มอักขระที่ต่างกัน กลุ่มภาษายุโรปตะวันตกใช้รหัส ISO-8859-1 สำหรับภาษาไทยใช้รหัส ISO-8859-11 (ตัวเลขมุมซ้ายแสดงรหัสอักขระในรูปเลขฐานสิบหก)

A0	A1 ¡	A2 ¢	A3 £	A4 ¤	A5 ¥	A6 ¦	A7 §	A8 ¨	A9 ©	AA ª	AB «	AC ¬	AD	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ¹	BA º	BB »	BC ¼	BD ½	BE ¾	BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
D0 Ð	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý	DE Þ	DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
F0 ð	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý	FE þ	FF ÿ

ตาราง 9.3 ตัวอักขระที่เพิ่มเติมมาใน ISO-8859-1 หรือ Latin-1

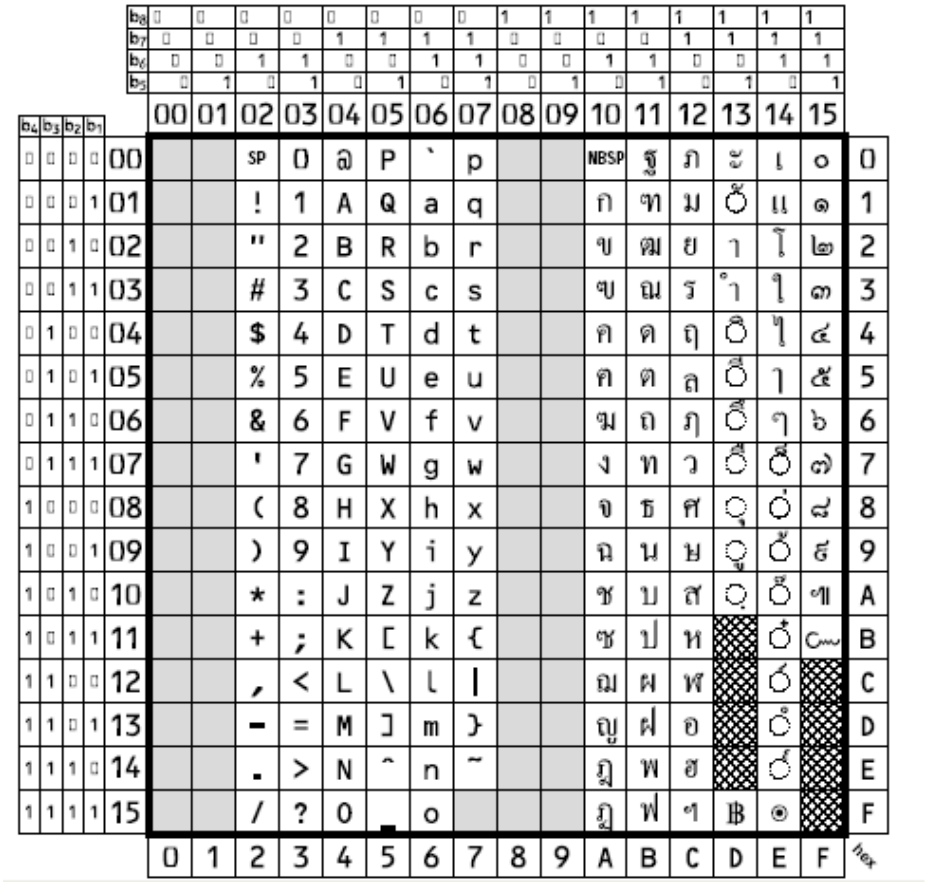
ISO-8859-x	Name	Year	Languages covered
1	Latin-1	1987	Western European languages
2	Latin-2	1987	East European languages
3	Latin-3	1988	Southern European languages
4	Latin-4	1988	Northern European languages
5	Latin/Cyrillic	1988	Russian, Bulgarian, Ukrainian, etc.
6	Latin/Arabic	1987	Arabic
7	Latin/Greek	1987	Greek
8	Latin/Hebrew	1988	Hebrew
9	Latin-5	1989	Turkish (Replaces Latin-3)
10	Latin-6		Northern European languages (Unifies Latin-1 and Latin-4)
11	Latin/Thai		Thai
12	Currently unassigned		May be used in future for Indian or Vietnamese
13	Latin-7		Baltic languages (Replaces Latin-4 and supplements Latin-6)
14	Latin-8		Celtic characters
15	Latin-9	1998	Western European languages (Replaces Latin-1 and adds the euro symbol plus a few missing French and Finnish characters)
16	Latin-10		Eastern European languages (Replaces Latin-2 and adds the euro symbol plus a few missing Romanian characters

ตาราง 9.4 รหัสตัวอักขระ ISO-8859 แบบต่างๆ

2.2 ตัวอักขระภาษาไทย

ในระยะแรกที่มีการเพิ่มเติมภาษาไทยเข้าไปในรหัสตัวอักษรของคอมพิวเตอร์ ก็ได้อาศัยรหัสในช่วง 128-255 ที่ไม่ได้ใช้สำหรับภาษาอังกฤษมาใช้กำหนดตัวอักษรต่างๆ ของภาษาไทย แต่ในระยะแรกนั้น การกำหนดรหัสอักขระ

ไทยก็เป็นไปแบบต่างคนต่างกำหนด ผู้ที่พัฒนาโปรแกรมพิมพ์งานทั้งหลายในยุคแรก เช่น ราชวิถีเวิร์ด เวิร์ดของเกษตร จึงใช้รหัสแทนอักขระไทยแตกต่างกัน ส่งผลให้ไม่สามารถนำข้อมูลที่พิมพ์จากโปรแกรมหนึ่งไปใช้กับอีกโปรแกรมหนึ่งได้ในทันที จำเป็นต้องทำการแปลงรหัสไปมา ต่อมาสำนักงานมาตรฐานอุตสาหกรรมจึงได้กำหนดมาตรฐานรหัสสำหรับอักขระไทยในปี พ.ศ.2529 เรียกว่ามาตรฐาน TIS 620 และได้ปรับปรุงเพิ่มเติมในปี พ.ศ. 2533 (ค.ศ.1990) ซึ่งมาตรฐาน ISO-8859-11 ที่ใช้สำหรับภาษาไทย(กำหนดในปีค.ศ. 1999) ก็ได้นำมาตรฐาน TIS 620-2533 นี้มาใช้เป็นหลัก ต่างกันแต่เพียงว่าในมาตรฐาน ISO-8859-11 กำหนดให้รหัส A0 เป็น non-breaking space คือช่องไฟที่ไม่ต้องปัดขึ้นบรรทัดใหม่ ส่วน TIS 620 ไม่ใช้รหัสนี้สำหรับตัวอักขระ) นอกจากตารางรหัสทั้งสองแล้ว สำหรับภาษาไทยยังมีการใช้ตารางรหัส Windows-874 สำหรับตัวอักขระภาษาไทยด้วย ซึ่งเป็นรหัสที่บริษัทไมโครซอฟต์กำหนดขึ้น แต่ผู้ใช้โดยทั่วไปจะไม่รู้สึกว่ามีความแตกต่างกันอย่างไร เพราะตัวอักขระไทยที่ใช้กันนั้นถูกกำหนดไว้เหมือนกัน ส่วนที่แตกต่างไปคือ ใน Windows-874 มีการกำหนดรหัส 80-9F ใช้ในขณะที่ TIS 620 ไม่ได้กำหนดรหัสส่วนนี้ให้ใช้แทนอักขระใด Windows-874, ISO-8859-11 จึงเป็นเหมือนส่วนขยาย (superset) ของมาตรฐาน TIS 620 คือกำหนดไว้มากกว่า



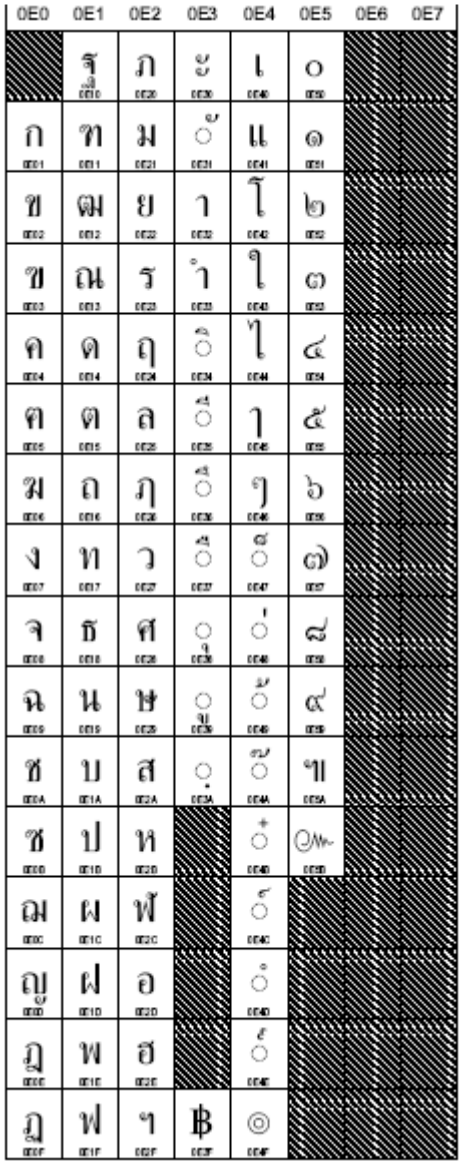
b_4 b_3 b_2 b_1	b_8 b_7 b_6 b_5	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	
0 0 0 0		00			SP	0	@	P	`	p			NBSP	ฐ	ภ	ะ	เ	๐	0
0 0 0 1		01			!	1	A	Q	a	q			ก	ฑ	ม	ั	แ	๑	1
0 0 1 0		02			"	2	B	R	b	r			ข	ฒ	ย	า	โ	๒	2
0 0 1 1		03			#	3	C	S	c	s			ฃ	ณ	ร	ำ	ใ	๓	3
0 1 0 0		04			$	4	D	T	d	t			ค	ด	ฤ	ิ	ไ	๔	4
0 1 0 1		05			%	5	E	U	e	u			ฅ	ต	ล	ี	ๅ	๕	5
0 1 1 0		06			&	6	F	V	f	v			ฆ	ถ	ฦ	ึ	ๆ	๖	6
0 1 1 1		07			'	7	G	W	g	w			ง	ท	ว	ื	็	๗	7
1 0 0 0		08			(	8	H	X	h	x			จ	ธ	ศ	ุ	่	๘	8
1 0 0 1		09			)	9	I	Y	i	y			ฉ	น	ษ	ู	้	๙	9
1 0 1 0		10			*	:	J	Z	j	z			ช	บ	ส	ฺ	๊	๚	A
1 0 1 1		11			+	;	K	[	k	{			ซ	ป	ห		๋	๛	B
1 1 0 0		12			,	<	L	\	l	|			ฌ	ผ	ฬ		์		C
1 1 0 1		13			-	=	M	]	m	}			ญ	ฝ	อ		๎		D
1 1 1 0		14			.	>	N	^	n	~			ฎ	พ	ฮ		๏		E
1 1 1 1		15			/	?	O	_	o				ฏ	ฟ	ฯ	฿	๑		F
		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	hex

ตาราง 9.5 รหัสอักขระของ ISO 8859-11 นับแบบเลขฐานสิบหก

2.3 รหัสยูนิโค้ด

รหัสแบบที่กำหนด 1 byte หรือ 8 บิตสำหรับ 1 ตัวอักษรนั้นมีความเหมาะสมและเพียงพอสำหรับภาษาที่ใช้ตัวอักขระแต่ละตัวแทนเสียงในภาษา (alphabetical language) แต่สำหรับภาษาที่เขียนโดยใช้ตัวอักขระแทนพยางค์หรือหน่วยคำ เช่น ภาษาญี่ปุ่น ภาษาจีน จะมีจำนวนอักขระที่เขียนเป็นจำนวน

มากเกินกว่าที่จะแทนด้วย 1 byte ได้ (ซึ่งแทนได้เพียง 256 แบบของตัวอักขระ) อีกทั้งเมื่อมีการใช้รหัสอักขระที่แตกต่างกัน ก็จะมีผลต่อการย้ายข้อมูลข้ามภาษา เช่น รหัสที่แทนตัวอักขระ ก ที่ใช้ภาษาไทยจะไปตรงกับรหัสตัวอักขระ ¡ ในอีกตารางหนึ่ง เป็นต้น จึงมีความพยายามแก้ปัญหาให้มีรหัสเดียวที่ใช้ได้กับอักขระทุกภาษา ซึ่งเป็นที่มาของการพัฒนารหัส Unicode ขึ้นมาตั้งแต่ปีค.ศ.1991 (Unicode 1.0) โดยที่รหัสตัวอักขระ 256 ตัวแรกนั้นจะเหมือนกับรหัสของ ISO-8859 ปัจจุบันรหัสนี้ได้พัฒนามาถึง Unicode 5.0 (ปีค.ศ.2006) หลักการของ Unicode คือแต่ละตัวอักขระที่ใช้ในภาษาต่างๆ จะมีรหัสเฉพาะของตน แต่รูปแบบของการเข้ารหัสแบบ Unicode นั้นมีได้หลายแบบ เป็นการแปลงรหัส Unicode ที่กำหนดหรือที่เรียกว่า Unicode transformation format (UTF) เป็นชุดของ byte เฉพาะสำหรับแต่ละอักขระ (unique byte sequence) ในแบบที่ง่ายสุดคือใช้จำนวนบิตคงที่ (เหมือนการแทนรหัสแบบอื่นที่กล่าวมา) จะใช้จำนวน 32 บิตสำหรับแต่ละอักขระ (4 bytes/character) (เรียกว่า UTF-32) แต่จะเสียเนื้อที่เก็บข้อมูลมากไม่ประหยัด และเพราะตัวอักษรต่างๆที่ใช้บ่อยๆ ก็มักเป็นเหมือนรหัส ISO-8859 ซึ่งใช้เพียง 8 บิตสำหรับแต่ละตัวอักษรก็พอ จึงมีการคิดวิธีการเข้ารหัสที่ประหยัดที่เก็บมากขึ้น โดยจำนวนบิตของแต่ละอักขระไม่จำเป็นต้องคงที่ การเข้ารหัสแบบ UTF-16 กำหนดให้ตัวอักขระที่ใช้บ่อยๆ เก็บ 16 บิต ส่วนตัวอักขระที่ใช้ไม่บ่อยจะเก็บเป็น 32 บิต (2-4 bytes/character) ส่วนการเข้ารหัสแบบ UTF-8 จะยิ่งประหยัดมากขึ้น เพราะตัวอักขระที่ใช้บ่อยจะเก็บ 8 บิต ส่วนตัวอักขระอื่นๆ อาจเก็บ 16 หรือ 32 บิต (1-4 bytes/character) จึงทำให้ประหยัดที่เก็บข้อมูลมากขึ้น รหัสแบบ unicode นี้เป็นที่คาดการณ์ว่าจะเป็นมาตรฐานสำหรับการเก็บข้อมูลภาษาในคอมพิวเตอร์ในอนาคต



ตาราง 9.6 รหัสภาษาไทยแบบ Unicode 0E00-0E7F

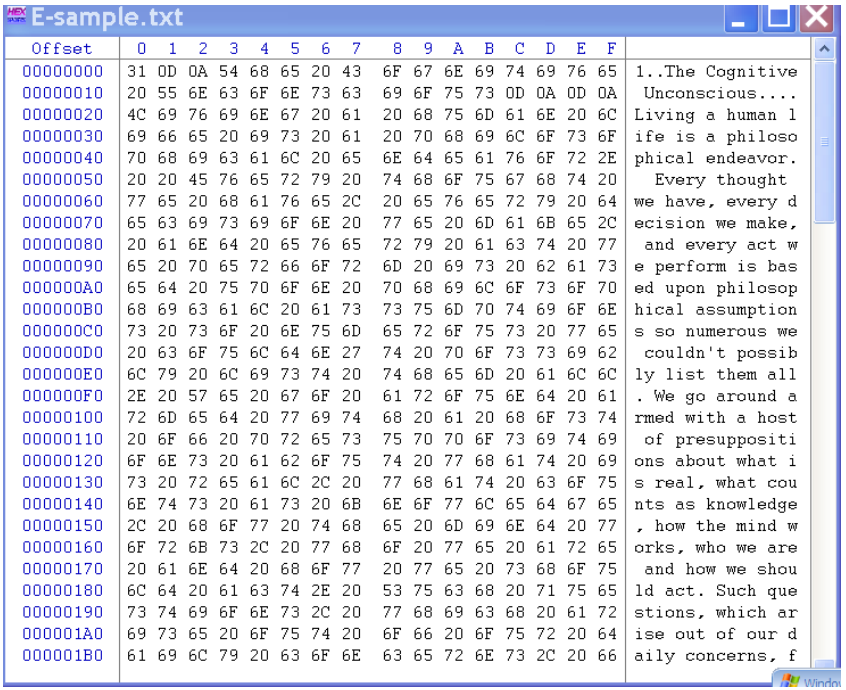
3. ไฟล์ข้อมูลแบบต่าง ๆ

ในการพิมพ์ข้อมูลภาษาเข้าไปเก็บในคอมพิวเตอร์นั้น ข้อความหรืออันที่จริงแต่ละตัวอักษรจะถูกแทนด้วยรหัสอักขระตามที่กล่าวมา ข้อความทั้งหมดจะถูกเก็บลงในไฟล์คอมพิวเตอร์ ถ้าเป็นการเก็บเฉพาะข้อมูลภาษาอย่างเดียวเราก็เรียกว่าเป็น plain text แต่ในการใช้คอมพิวเตอร์พิมพ์งานโดยทั่วไป นอกจากข้อความภาษาแล้ว ยังมีการจัดรูปแบบการแสดงผลหน้าจอหรือทางเครื่องพิมพ์ เช่น ทำตัวหนา ตัวเอียง ขีดเส้นใต้ เป็นต้น ไฟล์ข้อมูลจากโปรแกรมพิมพ์งานทั้งหลายอย่าง MS Word ก็จะเป็นลักษณะนี้ คือเก็บข้อมูลอื่นนอกเหนือจากข้อมูลภาษาด้วย ซึ่งโปรแกรมพิมพ์งานแต่ละโปรแกรมก็มีวิธีการจัดเก็บรูปแบบที่แตกต่างกัน เราจึงไม่สามารถนำข้อมูลที่พิมพ์จาก MS Word ไปเปิดดูในโปรแกรมอื่นอย่าง Word Perfect ได้ในทันที จะต้องมีการแปลงไฟล์ให้อยู่ในรูปที่โปรแกรมนั้นรู้จักก่อน ไฟล์ข้อมูลลักษณะนี้เรียกว่า formatted text

อย่างไรก็ตาม ไฟล์จากโปรแกรมพิมพ์งานเหล่านี้ เวลาไปเปิดบนเครื่องต่างชนิดก็อาจมีปัญหาได้รูปแบบไม่ตรงกับที่จัดไว้แต่เดิม จึงมีการสร้างไฟล์อีกประเภท ซึ่งเป็นไฟล์ที่คงรูปแบบตามที่ตั้งไว้แต่แรก เปรียบเหมือนเป็นการส่งภาพถ่ายจากที่หนึ่งไปอีกที่หนึ่ง ซึ่งจะไม่มีการผิดเพี้ยนของรูปแบบต่างๆ ไฟล์ประเภทนี้ เช่น ไฟล์ post script (.ps) ไฟล์ PDF (.pdf) ไฟล์ประเภทนี้บางครั้งก็จะไม่เห็นว่ามีข้อความที่พิมพ์จัดเก็บไว้ในไฟล์เลย

ไฟล์อีกประเภทหนึ่งที่ใช้กันแพร่หลายในปัจจุบัน คือไฟล์ที่เรียกว่า annotated text คือมีการเก็บข้อความภาษาพร้อมกับมีการกำกับข้อมูลว่าข้อความส่วนนั้นคืออะไร เรียกว่ามีการใส่แท็ก เช่น <title>Perl (Cookbook)</title> เป็นการแท็กว่า ชื่อ Perl (Cookbook) เป็นข้อมูลที่เป็น title ตัวอย่างของไฟล์ลักษณะนี้ เช่น ไฟล์ HTML หรือเว็บเพจทั้งหลาย

รูปที่ 9.1-9.3 แสดงให้เห็นถึงสิ่งที่เก็บอยู่ภายในไฟล์แต่ละประเภทในรูปที่เป็นรหัสเลขฐานสิบหก ในไฟล์ที่เป็นข้อความล้วน รูปที่ 9.1 เราจะเห็นว่าข้อความที่พิมพ์ไป คือ เลข 1 แล้วขึ้นบันทัดใหม่ “The Cognitive …” เมื่อเก็บในไฟล์จะเป็นรหัส 31 สำหรับ “1” รหัส 0D0A สำหรับขึ้นบรรทัดใหม่ รหัส 54 สำหรับ “T” รหัส 68 สำหรับ “h” รหัส 65 สำหรับ “e” เป็นลักษณะนี้ไปเรื่อยๆ


E-sample.txt

Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
00000000	31	0D	0A	54	68	65	20	43	6F	67	6E	69	74	69	76	65	1..The Cognitive
00000010	20	55	6E	63	6F	6E	73	63	69	6F	75	73	0D	0A	0D	0A	Unconscious....
00000020	4C	69	76	69	6E	67	20	61	20	68	75	6D	61	6E	20	6C	Living a human l
00000030	69	66	65	20	69	73	20	61	20	70	68	69	6C	6F	73	6F	ife is a philoso
00000040	70	68	69	63	61	6C	20	65	6E	64	65	61	76	6F	72	2E	phical endeavor.
00000050	20	20	45	76	65	72	79	20	74	68	6F	75	67	68	74	20	Every thought
00000060	77	65	20	68	61	76	65	2C	20	65	76	65	72	79	20	64	we have, every d
00000070	65	63	69	73	69	6F	6E	20	77	65	20	6D	61	6B	65	2C	ecision we make,
00000080	20	61	6E	64	20	65	76	65	72	79	20	61	63	74	20	77	and every act w
00000090	65	20	70	65	72	66	6F	72	6D	20	69	73	20	62	61	73	e perform is bas
000000A0	65	64	20	75	70	6F	6E	20	70	68	69	6C	6F	73	6F	70	ed upon philosop
000000B0	68	69	63	61	6C	20	61	73	73	75	6D	70	74	69	6F	6E	hical assumption
000000C0	73	20	73	6F	20	6E	75	6D	65	72	6F	75	73	20	77	65	s so numerous we
000000D0	20	63	6F	75	6C	64	6E	27	74	20	70	6F	73	73	69	62	couldn't possib
000000E0	6C	79	20	6C	69	73	74	20	74	68	65	6D	20	61	6C	6C	ly list them all
000000F0	2E	20	57	65	20	67	6F	20	61	72	6F	75	6E	64	20	61	. We go around a
00000100	72	6D	65	64	20	77	69	74	68	20	61	20	68	6F	73	74	rmed with a host
00000110	20	6F	66	20	70	72	65	73	75	70	70	6F	73	69	74	69	of presuppositi
00000120	6F	6E	73	20	61	62	6F	75	74	20	77	68	61	74	20	69	ons about what i
00000130	73	20	72	65	61	6C	2C	20	77	68	61	74	20	63	6F	75	s real, what cou
00000140	6E	74	73	20	61	73	20	6B	6E	6F	77	6C	65	64	67	65	nts as knowledge
00000150	2C	20	68	6F	77	20	74	68	65	20	6D	69	6E	64	20	77	, how the mind w
00000160	6F	72	6B	73	2C	20	77	68	6F	20	77	65	20	61	72	65	orks, who we are
00000170	20	61	6E	64	20	68	6F	77	20	77	65	20	73	68	6F	75	and how we shou
00000180	6C	64	20	61	63	74	2E	20	53	75	63	68	20	71	75	65	ld act. Such que
00000190	73	74	69	6F	6E	73	2C	20	77	68	69	63	68	20	61	72	stions, which ar
000001A0	69	73	65	20	6F	75	74	20	6F	66	20	6F	75	72	20	64	ise out of our d
000001B0	61	69	6C	79	20	63	6F	6E	63	65	72	6E	73	2C	20	66	aily concerns, f

รูป 9.1 ไฟล์ข้อมูลจัดเก็บเป็น plain text

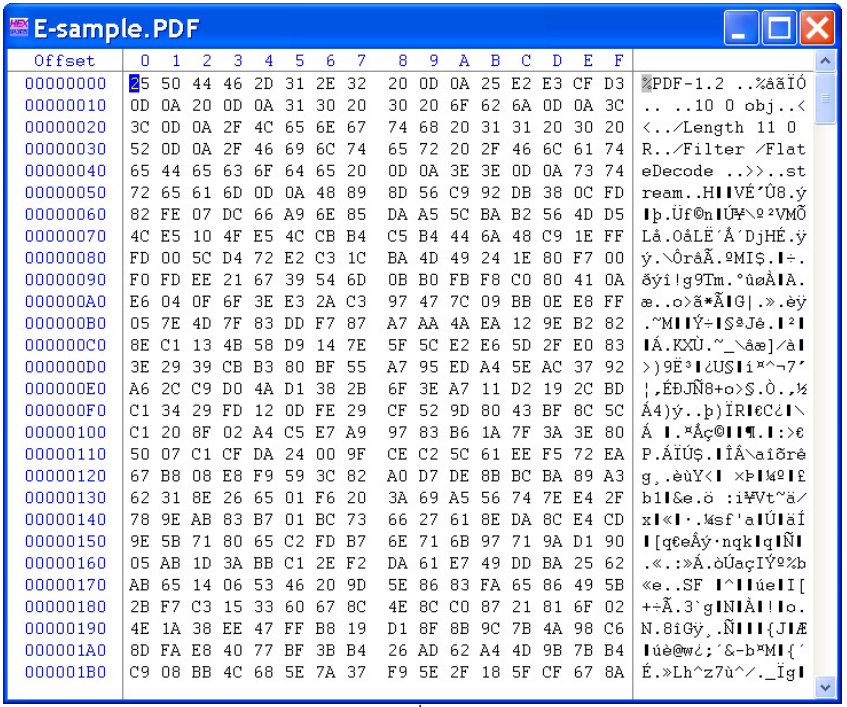
ส่วนรูปที่ 9.2 จะเห็นว่าสิ่งที่เก็บในไฟล์ MS Word จะมีการเก็บสิ่งอื่นๆไว้ในตอนต้นของไฟล์ด้วย ซึ่งเป็นรูปแบบเฉพาะของโปรแกรมที่กำหนดมาเพื่อใช้ในการจัดรูปแบบเอกสารในไฟล์นั้น แล้วจึงมีส่วนที่เป็นข้อความในตอนหลัง

Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
00000000	D0	CF	11	E0	A1	B1	1A	E1	00	00	00	00	00	00	00	00	ÐÏ.à¡±.á........
00000010	00	00	00	00	00	00	00	00	3E	00	03	00	FE	FF	09	00	........>...þÿ..
00000020	06	00	00	00	00	00	00	00	00	00	00	00	01	00	00	00	....
00000030	26	00	00	00	00	00	00	00	00	10	00	00	28	00	00	00	&...........(...
00000040	01	00	00	00	FE	FF	FF	FF	00	00	00	00	25	00	00	00	....þÿÿÿ....%...
00000050	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000060	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000070	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000080	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000090	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000000A0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000000B0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000000C0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000000D0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000000F0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000100	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000110	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000120	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000130	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000140	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000150	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000160	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000170	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000180	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
00000190	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000001A0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000001B0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
000001C0	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	FF	ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ

Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
000005C0	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	....
000005D0	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	....
000005E0	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	....
000005F0	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	....
00000600	31	0D	54	68	65	20	43	6F	67	6E	69	74	69	76	65	20	1.The Cognitive
00000610	55	6E	63	6F	6E	73	63	69	6F	75	73	0D	0D	4C	69	76	Unconscious..Liv
00000620	69	6E	67	20	61	20	68	75	6D	61	6E	20	6C	69	66	65	ing a human life
00000630	20	69	73	20	61	20	70	68	69	6C	6F	73	6F	70	68	69	is a philosophi
00000640	63	61	6C	20	65	6E	64	65	61	76	6F	72	2E	20	20	45	cal endeavor. E
00000650	76	65	72	79	20	74	68	6F	75	67	68	74	20	77	65	20	very thought we
00000660	68	61	76	65	2C	20	65	76	65	72	79	20	64	65	63	69	have, every deci
00000670	73	69	6F	6E	20	77	65	20	6D	61	6B	65	2C	20	61	6E	sion we make, an
00000680	64	20	65	76	65	72	79	20	61	63	74	20	77	65	20	70	d every act we p
00000690	65	72	66	6F	72	6D	20	69	73	20	62	61	73	65	64	20	erform is based
000006B0	61	6C	20	61	73	73	75	6D	70	74	69	6F	6E	73	20	73	al assumptions s
000006C0	6F	20	6E	75	6D	65	72	6F	75	73	20	77	65	20	63	6F	o numerous we co
000006D0	75	6C	64	6E	27	74	20	70	6F	73	73	69	62	6C	79	20	uldn't possibly
000006E0	6C	69	73	74	20	74	68	65	6D	20	61	6C	6C	2E	20	57	list them all. W
000006F0	65	20	67	6F	20	61	72	6F	75	6E	64	20	61	72	6D	65	e go around arme
00000700	64	20	77	69	74	68	20	61	20	68	6F	73	74	20	6F	66	d with a host of
00000710	20	70	72	65	73	75	70	70	6F	73	69	74	69	6F	6E	73	presuppositions
00000720	20	61	62	6F	75	74	20	77	68	61	74	20	69	73	20	72	about what is r
00000730	65	61	6C	2C	20	77	68	61	74	20	63	6F	75	6E	74	73	eal, what counts
00000740	20	61	73	20	6B	6E	6F	77	6C	65	64	67	65	2C	20	68	as knowledge, h
00000750	6F	77	20	74	68	65	20	6D	69	6E	64	20	77	6F	72	6B	ow the mind work
00000760	73	2C	20	77	68	6F	20	77	65	20	61	72	65	20	61	6E	s, who we are an
00000770	64	20	68	6F	77	20	77	65	20	73	68	6F	75	6C	64	20	d how we should
00000780	61	63	74	2E	20	53	75	63	68	20	71	75	65	73	74	69	act. Such questi

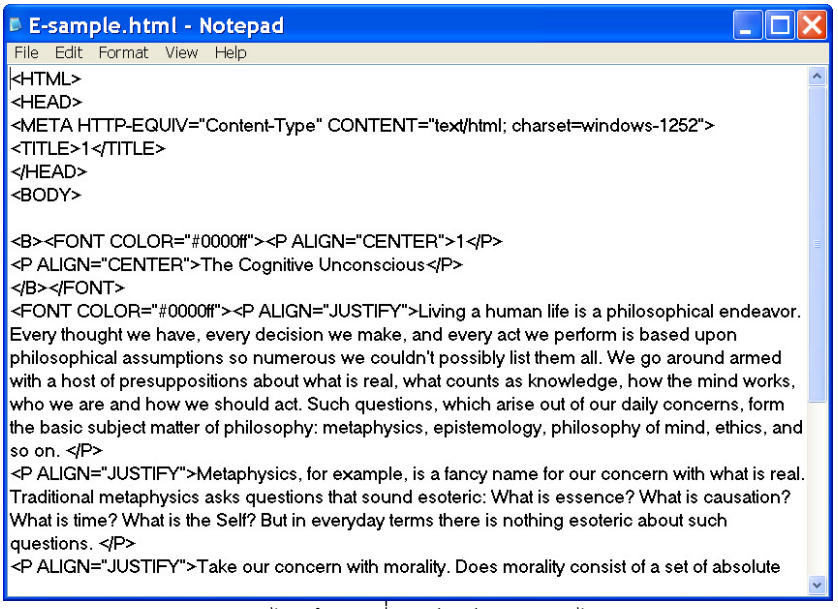
รูป 9.2 ไฟล์ข้อมูลที่จัดเก็บเป็น MS Word ไฟล์

ส่วนรูปที่ 9.3 ซึ่งเป็นไฟล์ pdf นั้นเราจะไม่เห็นว่ามีการเก็บข้อความที่พิมพ์ไว้ในไฟล์ด้วยรหัสอักขระใดใด แต่กลายเป็นการเก็บตามรูปแบบเฉพาะที่กำหนดโดย Acrobat เอง ในส่วนต้นไฟล์จะเห็นมีการใส่ว่าเป็น %PDF-1.2 หมายถึง pdf version 1.2



รูป 9.3 ไฟล์ข้อมูลที่จัดเก็บเป็น PDF

ส่วนรูปที่ 9.4 เป็นไฟล์ html ซึ่งเราจะเห็นข้อความตามปกติ แต่จะมีการใส่แท็กต่างๆ เช่น <head> <body> เป็นต้น



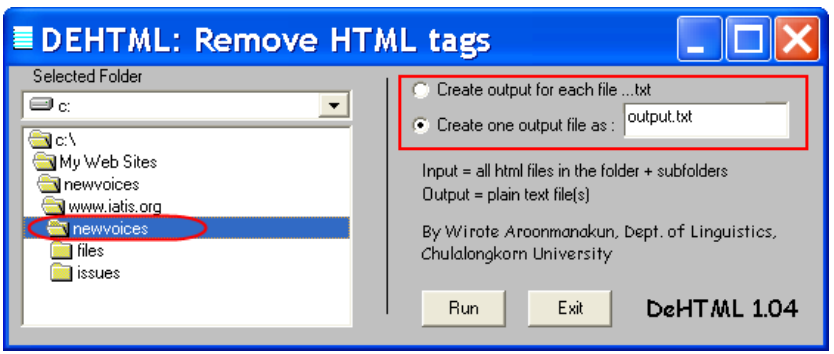
รูป 9.4 ไฟล์ข้อมูลที่จัดเก็บเป็น HTML ไฟล์

4. การแปลงไฟล์ข้อมูล

ข้อมูลพื้นฐานที่ใช้ในการสร้างคลังข้อมูล คือ ข้อมูลที่เป็นข้อความล้วน (plain text) ซึ่งหากจำเป็นต้องมีการกำกับข้อมูลหรือใส่แท็กต่างๆ ก็จะสามารถทำได้ในภายหลัง ไฟล์ข้อมูลที่ได้มาในรูปแบบต่างๆ จะมีวิธีในการแปลงเป็นข้อความล้วนต่างกันไป

ข้อมูล MS Word ข้อมูลที่อยู่ในรูป MS Word สามารถจัดเก็บใหม่เป็นไฟล์ข้อความล้วนได้ ด้วยการเลือก Save as และเลือกประเภทข้อมูลเป็น plain text เฉพาะส่วนที่เป็นข้อมูลภาษาจะถูกเก็บในไฟล์ใหม่นี้ ข้อมูลที่บอกการจัดรูปแบบ (format) จะหายไปหมด

ข้อมูล HTML สามารถแปลงเป็นไฟล์ข้อความล้วนได้ด้วยวิธีทำนองเดียวกัน โดยการใช้คำสั่ง Save Page/Frame as แล้วเลือก save เป็นไฟล์ text บางโปรแกรม เช่น Internet Explorer เวลา save จะให้เลือก encoding ของตัวอักษรได้ด้วย ในกรณีที่ต้องการทำเช่นนี้กับไฟล์จำนวนมากซึ่งคัดลอกมาไว้ที่เครื่องคอมพิวเตอร์แล้ว ก็สามารถทำได้โดยใช้โปรแกรม DeHTML ซึ่งใช้สำหรับถอดแท็กต่างๆ ออกให้เหลือเฉพาะข้อความเท่านั้น โปรแกรมนี้สามารดาว์นโหลดได้ที่ http://pioneer.chula.ac.th/~awirote/ เมื่อติดตั้งโปรแกรมแล้ว เวลาใช้งานให้เลือกไปที่ folder ที่เก็บข้อมูล html ที่ต้องการแปลง จากนั้นเลือกวิธีการแปลงได้สองแบบ แบบแรกคือแปลงแต่ละไฟล์เป็น .txt ภายใต้ชื่อเดิม แบบสองคือแปลงข้อมูลทั้งหมดรวมลงในไฟล์เดียวตามชื่อที่กำหนดในหน้าจอ โปรแกรมจะต้นหาไฟล์ .htm หรือ .html ทุกไฟล์ใน folder นั้นตลอดจน folder ย่อยทั้งหมดที่อยู่ใต้ folder ที่เลือก

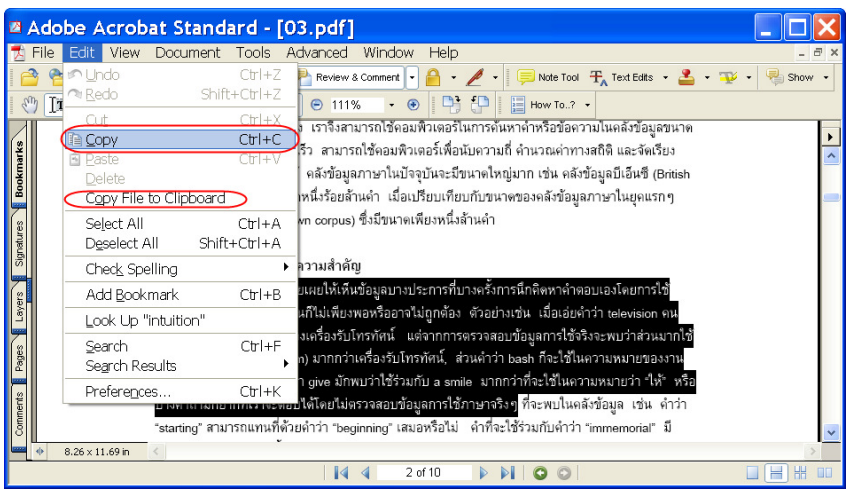


รูป 9.5 โปรแกรม DeHTML

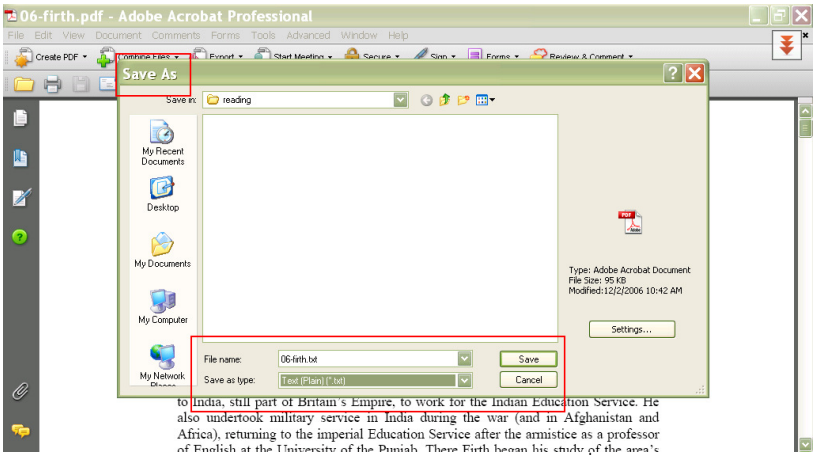
อย่างไรก็ตาม ข้อมูลที่แปลงจาก HTML นี้ บางครั้งก็มีข้อความบางส่วนที่ไม่ต้องการติดมาด้วยเนื่องจากหน้าโฮมเพ่จที่คัดลอกมานั้นมีการจัดเป็นเฟรมหรือมีแถบตัวเลือกต่างๆ อยู่ จึงอาจจำเป็นต้องนำมาแก้ไขลบข้อความที่ไม่ต้องการทิ้งเองในภายหลัง อีกวิธีหนึ่งที่ทำให้แน่ใจว่าจะคัดลอกเฉพาะส่วนของตัวบทที่

ต้องการ ก็คือการเลือกข้อความที่ต้องการด้วยตนเอง แล้วใช้คำสั่ง Edit – Copy ตามปกติ นำส่วนนั้นไป paste ลงในโปรแกรม Notepad

ข้อมูล pdf ข้อมูลที่ได้มาเป็น pdf หากเป็นไฟล์ที่สร้างจากโปรแกรมอื่นๆ ในคอมพิวเตอร์ เช่น โปรแกรม MS Word โดยทั่วไปจะสามารถแปลงเป็น text ได้ แต่ถ้าเป็นไฟล์ pdf ที่ได้จากการสแกนหน้ากระดาษเข้ามาเป็นภาพจะไม่สามารถแปลงข้อมูลเป็น text ได้ หากไฟล์นั้นสามารถแปลงข้อมูลได้จะสามารถเลือกข้อความที่ต้องการและสั่ง Edit – Copy ได้เพื่อคัดลอกเฉพาะส่วนที่เลือกไว้หรือจะใช้ Edit – Copy File to Clipboard เพื่อคัดลอกทั้งไฟล์ จากนั้นจึงแปะ (paste) ลงในโปรแกรม Notepad หรือจะใช้คำสั่ง File-Save as แล้วเลือกประเภทไฟล์ให้เป็น plain text หรือ .txt คำสั่งต่างๆ นี้อาจต่างกันไปตามเวอร์ชั่นของโปรแกรม Acrobat ที่ใช้ แต่หลักการจะเป็นในทำนองเดียวกันนี้

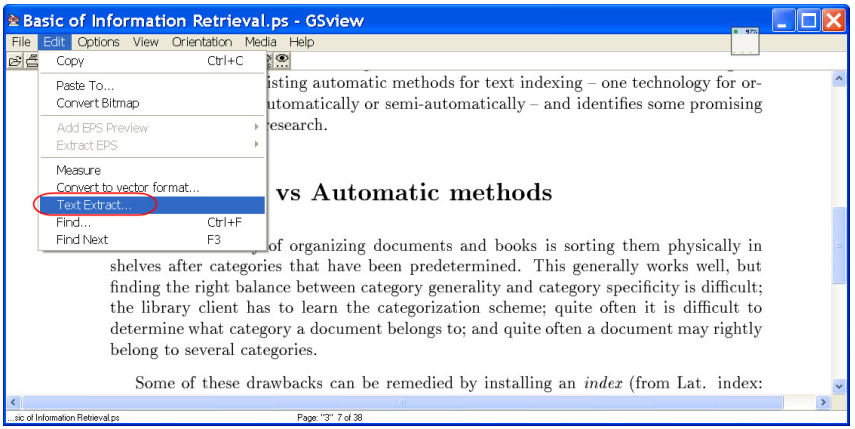


รูป 9.6 การแปลงข้อมูล pdf เป็นข้อความล้วนด้วยวิธี Edit - Copy



รูป 9.7 การแปลงข้อมูล pdf เป็นข้อความล้วนด้วยคำสั่ง File – Save As

ข้อมูล ps เป็นข้อมูลซึ่งสร้างขึ้นสำหรับส่งให้เครื่องพิมพ์ที่เป็น postscript printer มีลักษณะคล้ายกับไฟล์ pdf คือทั้งข้อความและการจัดรูปแบบจะอยู่คงที่ไม่ว่าจะเปิดที่เครื่องใดก็ตาม ไฟล์ .ps นี้สามารถเรียกดูหน้าจอได้ แต่ต้องมีการติดตั้งโปรแกรม Ghostscript และโปรแกรม GSview เสียก่อน โปรแกรมทั้งสองนี้สามารถดาว์นโหลดได้ที่ http://pages.cs.wisc.edu/~ghost/ หลังจากติดตั้งโปรแกรมทั้งสองแล้ว เมื่อเปิดไฟล์ .ps ด้วยโปรแกรม GSview แล้วจะสามารถแปลงข้อมูลเป็นไฟล์ text ได้โดยการเรียก Edit – Text Extract ซึ่งจะเลือกต่อได้ว่าต้องการแปลงข้อมูลในหน้าใดบ้าง

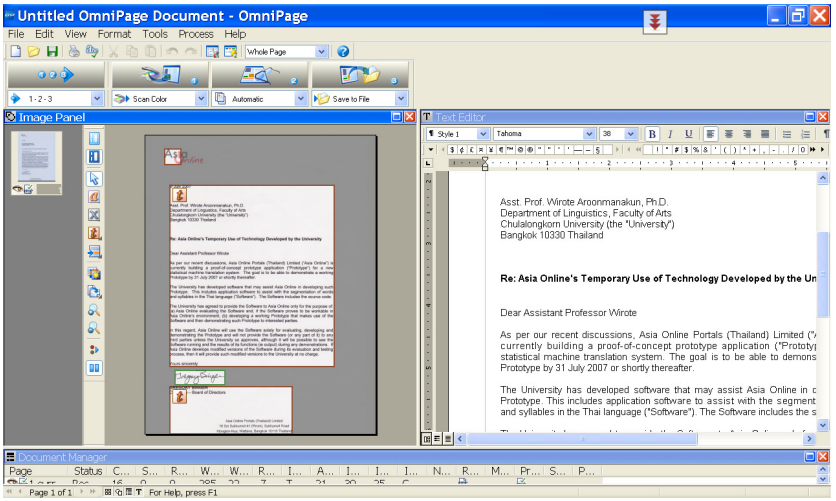


รูป 9.8 การแปลงไฟล์ .ps เป็นข้อความล้วน

5. การใช้โปรแกรมรู้จำตัวอักษรและสแกนเนอร์

ข้อมูลที่อยู่ในกระดาษสามารถนำเข้าในคอมพิวเตอร์ได้ หากมีเครื่องสแกนเนอร์และโปรแกรมรู้จำตัวอักษรในภาษานั้นที่มีประสิทธิภาพ หลักการคือใช้เครื่องสแกนเนอร์เพื่ออ่านข้อมูลเข้ามาเป็นภาพเหมือนเป็นการใช้เครื่องถ่ายเอกสารปกติ จากนั้นโปรแกรมรู้จำตัวอักษรจะมองข้อความนั้นและดูรูปร่างที่เห็นเพื่อคาดเดาว่าน่าจะเป็นอักษรใดในภาษานั้น วิธีการนี้จะได้ผลดีในภาษาที่มีโปรแกรมรู้จำตัวอักษรที่มีประสิทธิภาพสูง เช่น ภาษาอังกฤษ โปรแกรมรู้จำตัวอักษรนี้มักแถมมาพร้อมกับเครื่องสแกนเนอร์ เช่น โปรแกรม OmniPage SE ซึ่งโดยปกติ ทุกอย่างจะทำงานโดยอัตโนมัติได้ เมื่อสั่งโปรแกรมให้สแกนหน้ากระดาษ โปรแกรมจะสแกนภาพและกำหนดขอบเขตข้อความ รูปภาพ และวิเคราะห์รู้จำส่วนที่เป็นข้อความ ในขั้นตอนนี้ หากไม่แน่ใจว่าตัวอักษรที่พิจารณาเป็นตัวใดก็จะขึ้นหน้าจอมาถามผู้ใช้ได้ เมื่อเสร็จสิ้นแล้วก็จะสามารถบันทึกเป็นไฟล์ในรูปแบบต่างๆ เช่น MS Word, rtf, pdf, txt เป็นต้น หากเป็นการใช้เพื่อเก็บ

ข้อมูลเข้าในคลังข้อมูลภาษา การเก็บเป็นข้อความล้วนหรือ txt ก็เพียงพอแก่ความต้องการ



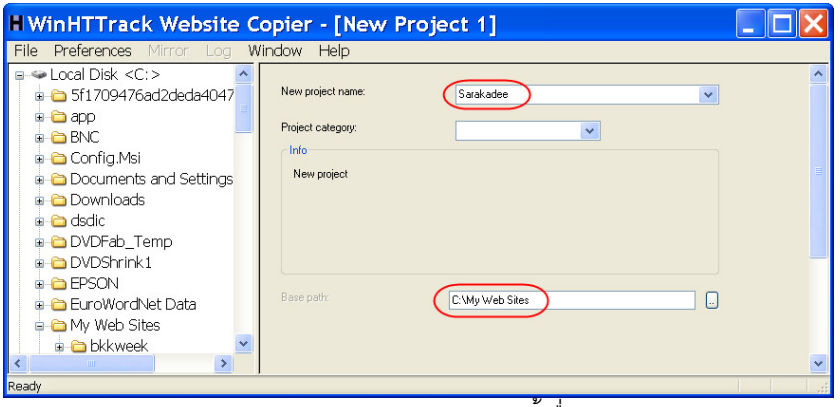
รูป 9.9 การใช้โปรแกรม OmniPage สแกนและรู้จำตัวอักษร

6. การใช้โปรแกรม HTTrack

โปรแกรม HTTrack เป็นโปรแกรมสำหรับใช้ดาวน์โหลดข้อมูลเว็บเพจจำนวนมากได้ โดยผู้ใช้สามารถกำหนดลักษณะการคัดลอกข้อมูลที่ต้องการได้ โปรแกรมนี้สามารถนำมาใช้ได้ฟรีโดยเข้าไปที่เว็บไซต์ http://www.httrack.com/ โปรแกรมนี้เขียนขึ้นมาเพื่อช่วยในการคัดลอกข้อมูลจากเว็บไซต์ที่ต้องการมาเก็บไว้ที่เครื่องของเราเอง ซึ่งจะทำให้เราสามารถเปิดอ่านข้อมูลต่างๆ ในภายหลังได้โดยไม่ต้องต่อกับอินเทอร์เน็ต ผลพลอยได้ของโปรแกรมนี้ในการสร้างคลังข้อมูลภาษา คือ เราสามารถโหลดเพจจำนวนมากจากเว็บไซต์ที่ต้องการมาได้ในคราวเดียวโดยอัตโนมัติ จากนั้นจึงสามารถแปลงไฟล์ทั้งหลายที่ได้มาเป็นไฟล์ .txt

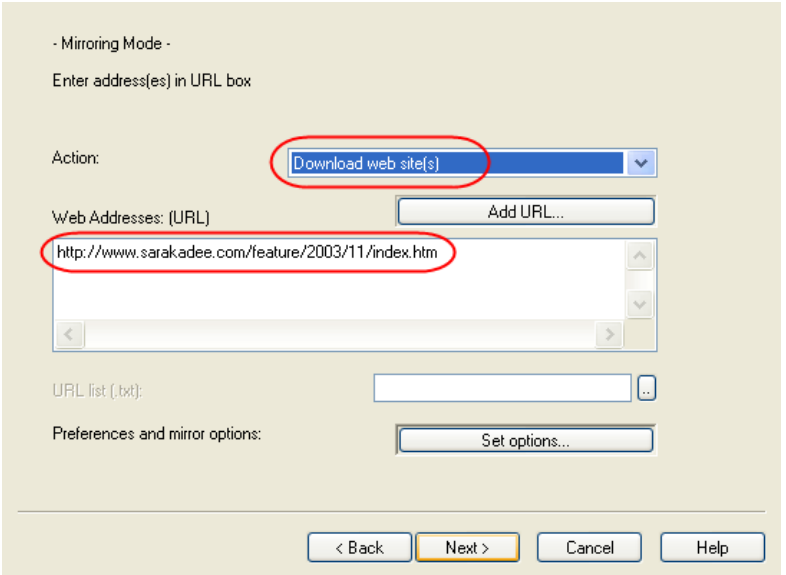
ด้วยการใช้โปรแกรม DeHTML ขั้นตอนการใช้โปรแกรม HTTrack หลังจากติดตั้งแล้วเป็นดังนี้

1. ตั้งชื่อโครงการ (project name) หรือเลือกโครงการที่ทำไว้ก่อนแล้ว ข้อมูลต่างๆ ที่โหลดมาได้จะเก็บไว้ที่ c:\My Web Sites\ แล้วตามด้วยชื่อที่ตั้งนี้



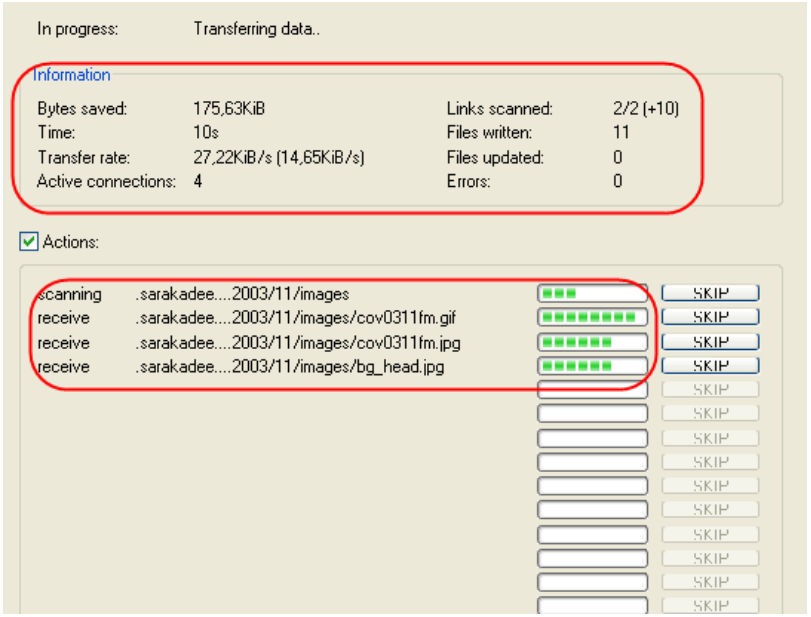
รูป 9.10 เปิดโปรแกรม HTTrack และตั้งชื่อโครงการ

2. เลือกสิ่งที่ต้องการทำ เช่น เริ่มต้นโหลดใหม่ (Download web sites) โหลดต่อจากครั้งก่อน (Continue interrupted download) ปรับปรุงข้อมูลที่โหลด (Updating existing download) เป็นต้น โดยเราจะต้องระบุ URL เริ่มต้นที่ต้องการ เช่น ถ้าต้องการโหลดข้อมูลจากเว็บไซ้ต์ สารคดี ฉบับเดือนพฤศจิกายน 2546 ก็ให้พิมพ์ URL นั้นในช่อง Web Address

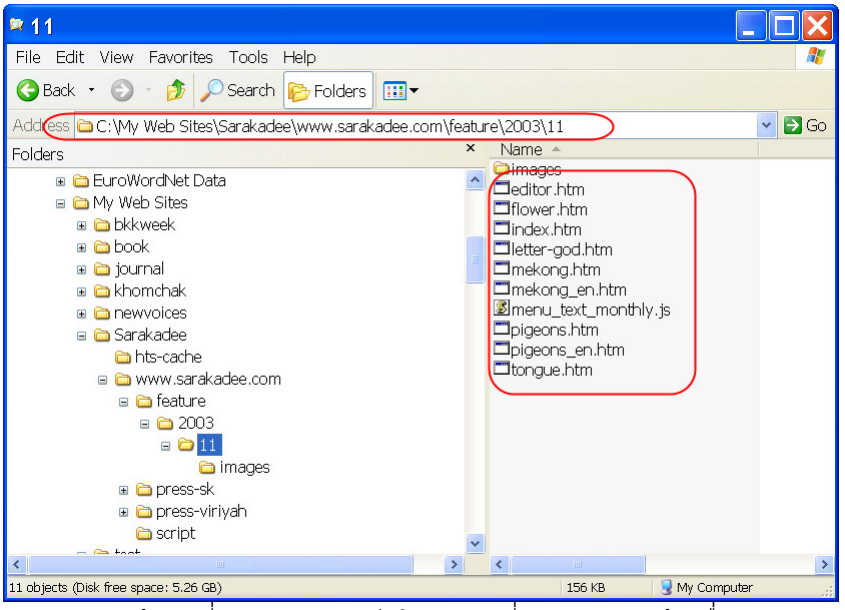


รูป 9.11 กำหนด url ที่ต้องการไปและเลือกโหมดการดาวน์โหลด

หลังจากสั่งให้ทำงานแล้ว จะเห็นหน้าจอบอกสถานะของการโหลดข้อมูลเข้ามาทั้งจำนวนไฟล์ที่เก็บลงเครื่องแล้ว ไฟล์ที่กำลังโหลดอยู่ เป็นต้น ซึ่งข้อมูลทั้งหมดหลังจากโหลดเสร็จแล้วจะถูกเก็บไว้ที่โฟลเดอร์ c:\My Web Sites\Sarakadee\ ตามชื่อโครงการที่ตั้งไว้ และจะมีการรตั้งชื่อโฟลเดอร์ย่อยๆ ตามลักษณะโครงสร้างที่เก็บในเว็บไซ้ต์นั้น

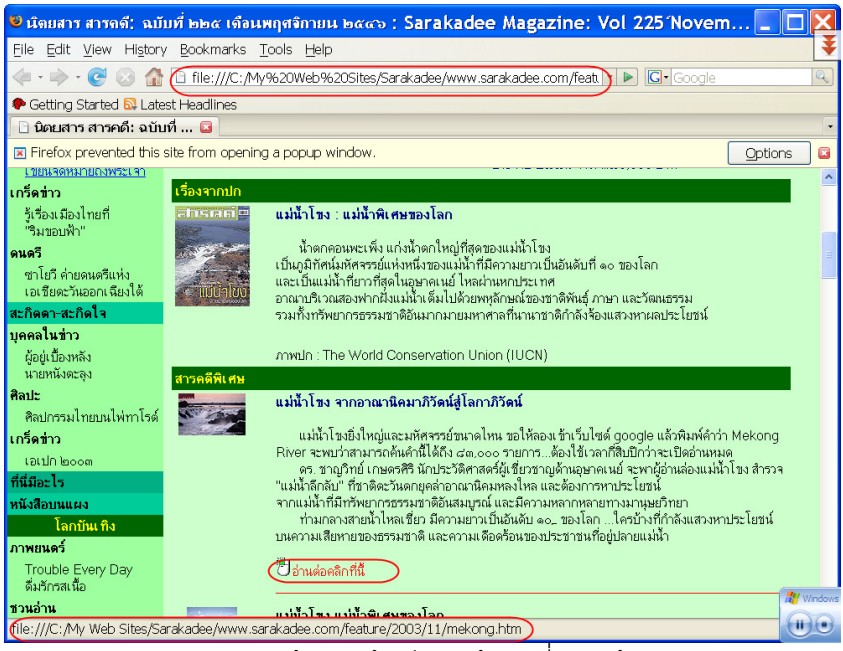


รูป 9.12 หน้าจอแสดงสถานะการโหลดข้อมูล



รูป 9.13 ข้อมูลที่โหลดมาถูกจัดเก็บใน folder ที่ระบุตามโครงสร้างชื่อ url

เมื่อเรียกดูข้อมูลในไฟล์ก็จะเห็นเหมือนที่ปรากฏบนอินเทอร์เน็ต เพียงแต่ข้อมูลเหล่านั้นนำมาจากที่เก็บไว้ในเครื่องของเราเอง ดังจะเห็นจาก address ที่อ้างถึง file:///c:/My Web Sites/Sarakadee ...

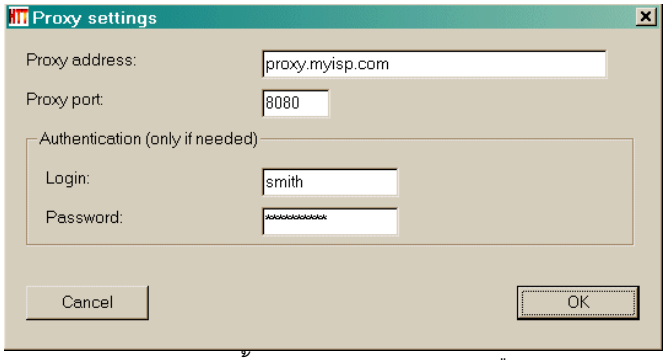


รูป 9.14 แสดงข้อมูลหน้าเว็บจากข้อมูลที่โหลดเข้ามา

6.1 ตัวเลือกต่าง ๆในโปรแกรม HTTrack

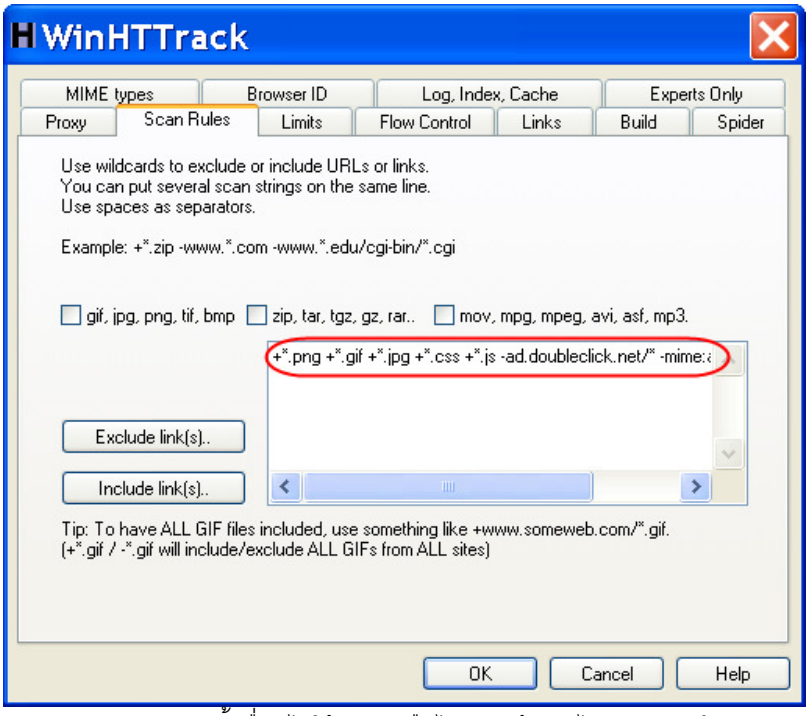
เราสามารถกำหนดค่าตัวเลือกต่างๆ ได้ใน Set options ซึ่งจะปรากฏเป็นแท็ปต่างๆ ให้เลือก ดังนี้

Proxy ใช้กำหนดค่า proxy ของเครือข่ายที่ใช้บริการอยู่ โดยเลือกที่ configure แล้วใส่ค่าที่ต้องการ ซึ่งต้องหาข้อมูลจากผู้ให้บริการว่าจะต้องกำหนดค่านี้หรือไม่ หากเครือข่ายไม่ได้ใช้ proxy ก็ไม่ต้องกำหนดค่าใดใด



รูป 9.15 การตั้งค่า proxy สำหรับบางเครือข่าย

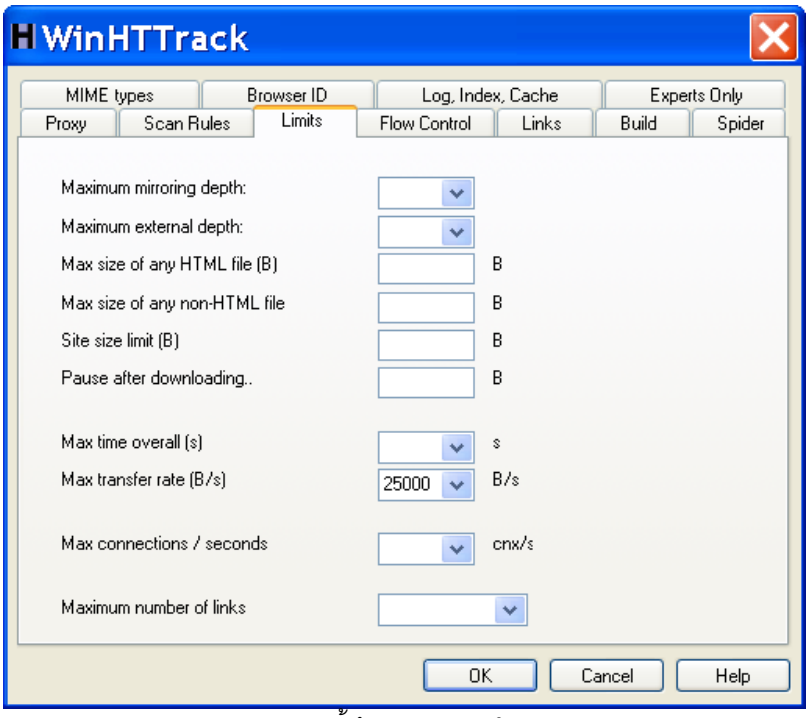
Scan Rules ใช้กำหนดเพิ่มเติมว่าให้โหลดไฟล์ประเภทใดอีกบ้าง และไม่โหลดไฟล์ประเภทใด กำหนดโดยใส่เครื่องหมาย + หรือ – หน้าชื่อกลุ่มไฟล์ที่ระบุ การระบุชื่อกลุ่มไฟล์ใช้อักขระพิเศษอย่าง * แทนตัวอักษรอะไรก็ได้ +*.zip จึงหมายถึงเก็บทุกไฟล์ที่สกุล .zip เป็นต้น



รูป 9.16 การตั้งเงื่อนไขให้โหลดหรือไม่โหลดข้อมูลไฟล์ประเภทใด

Limits ใช้จำกัดการโหลดข้อมูลลักษณะต่างๆ Maximum mirroring depth เป็นตัวเลือกที่ใช้บ่อย คือ ใช้กำหนดว่าจากหน้าเพ่จเริ่มต้นจะให้ไล่ลงลึกไปอีกเท่าไหร่ หากเป็น 1 คือเฉพาะหน้าแรกนั้น หากเป็น 2 คือหน้านั้นและหน้าอื่นๆ ที่ไล่ต่อจากหน้าแรกนั้น หากไม่กำหนดคือไล่ไปจนกว่าจะหมด ส่วน Maximum external depth ใช้กำหนดความลึกที่ต่อไปสำหรับไซ้ต์ข้างนอก(ที่ไม่ใช่ไซ้ต์ที่กำหนดในหน้าแรก) นอกจากนั้นยังกำหนดขนาดสูงสุดของไฟล์ที่จะโหลดเข้ามาได้ หรือสั่งให้หยุดชั่วคราวเมื่อโหลดข้อมูลถึงขนาดที่กำหนด กำหนดเวลามากสุดที่จะใช้ (Max time overall) จำกัดอัตราการรับข้อมูลไม่ให้เกินที่กำหนด (Max

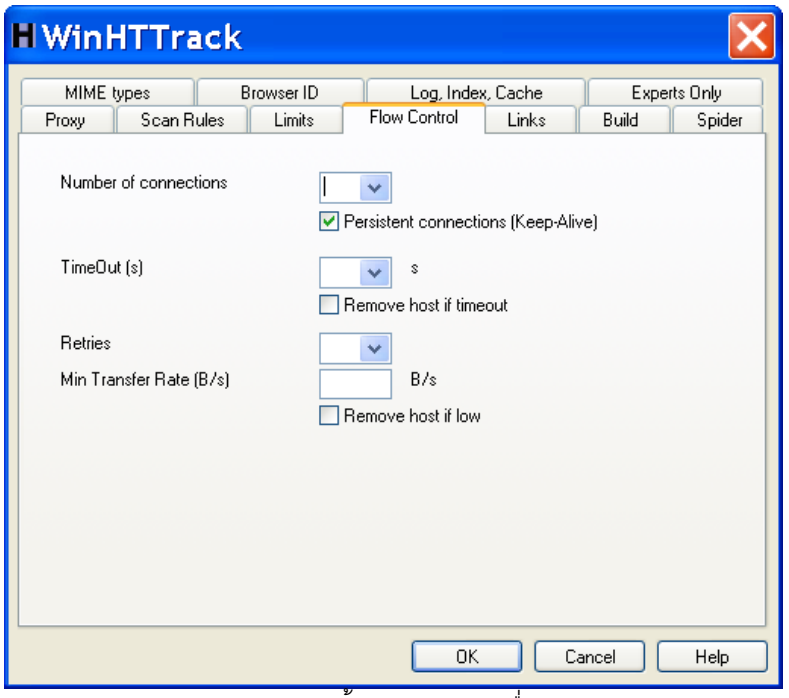
transfer rate) กำหนดจำนวนการเชื่อมต่อกับเครื่องแม่ข่าย (server) ว่าในหนึ่งวินาทีมีได้เท่าใด (Max connections/seconds) ถ้าไม่กำหนดจะเป็น 10 กำหนดจำนวนการเชื่อมโยงสูงสุด (link) ที่จะวิเคราะห์ว่าจะโหลดหรือไม่โหลดไฟล์นั้นๆ ค่าที่ให้ไว้คือ 100,000 ตัวเลือกต่างๆที่กล่าวมาโดยปกติจะไม่ใช้และไม่ควรเปลี่ยนแปลงอะไร ควรตั้งเฉพาะค่า mirroring depth ก็เพียงพอแล้ว



รูป 9.17 การตั้งข้อจำกัดต่างๆในการโหลด

Flow Control ใช้ควบคุมการดึงข้อมูลผ่านการเชื่อมโยง Number of connections คือจำนวนการเชื่อมต่อพร้อมๆกันที่ยอมให้เกิดได้ ถ้าไฟล์ข้อมูลที่โหลดส่วนใหญ่มีขนาดใหญ่ก็ควรกำหนดจำนวนเป็น 1 หรือ 2 ถ้าเป็นไฟล์ทั่วๆไป

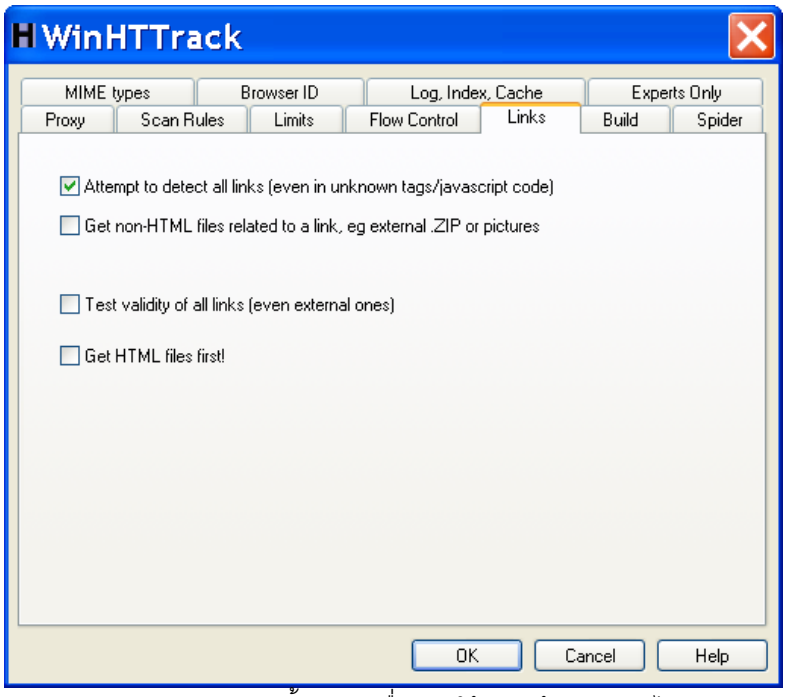
ก็อาจกำหนดเป็น 8 ได้ ยิ่งกำหนดมากจะยิ่งไปเพิ่มโหลดงานให้กับเครื่องแม่ข่ายได้ ส่วนตัวเลือกอื่นๆ นั้นไม่ค่อยได้ใช้ TimeOut กำหนดเวลาว่าให้คอยการตอบรับการเชื่อมโยงนานเท่าใด Retries ให้ลองเชื่อมต่อได้กี่ครั้งเมื่อเชื่อมต่อไม่ได้ Min Transfer Rate กำหนดว่าถ้าดึงข้อมูลได้ด้วยอัตราต่ำหว่าที่กำหนดก็ให้เลิกสนใจไฟล์นั้นๆ



รูป 9.18 การตั้งค่าจำนวนการเชื่อมต่อ

Links ใช้กำหนดการวิเคราะห์การเชื่อม โดยปกติจะเลือก Attempt to detect all links ไว้เพื่อให้โปรแกรมมองหา link ทั้งหลายในเพ่จนั้นแม้ว่าจะอยู่ในส่วนที่เป็นสคริปต์ ส่วน Get non-HTML files related to a link จะเลือกเมื่อเรา

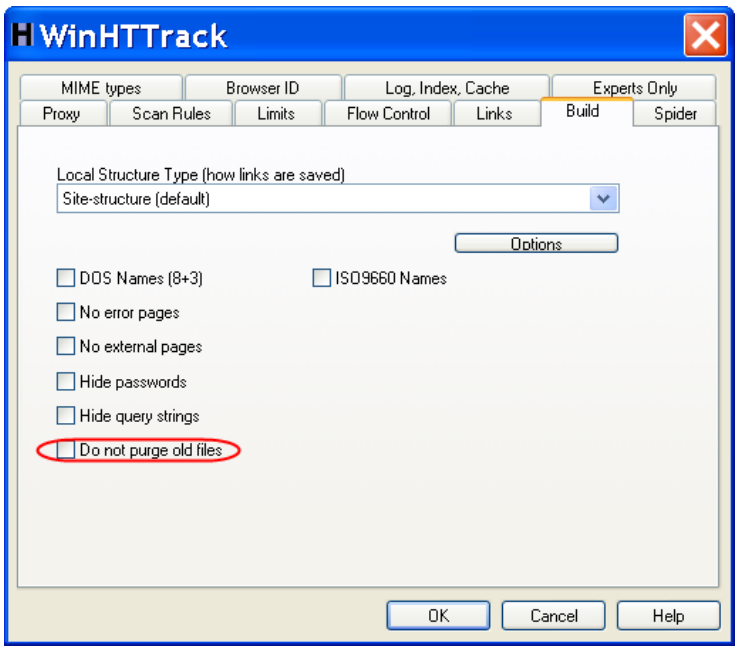
ต้องการให้โหลดไฟล์อื่นๆที่ไม่ใช่ html ไฟล์ด้วยหากมี link ไปถึง Test validity of all links ใช้เมื่อต้องการตรวจสอบ link ทั้งหมดที่จะต่อออกไปว่าใช้ได้หรือไม่ link ไหนไปยังเพ่จที่ไม่มีแล้วจะแสดงไว้ในรายการไฟล์ log ซึ่งจะเป็นประโยชน์สำหรับตรวจสอบไซ้ต์นั้นว่ามี link ที่ใช้ไม่ได้หรือไม่ ส่วน Get HTML files first ใช้บังคับให้เลือกโหลดไฟล์ html ก่อนไฟล์ประเภทอื่น



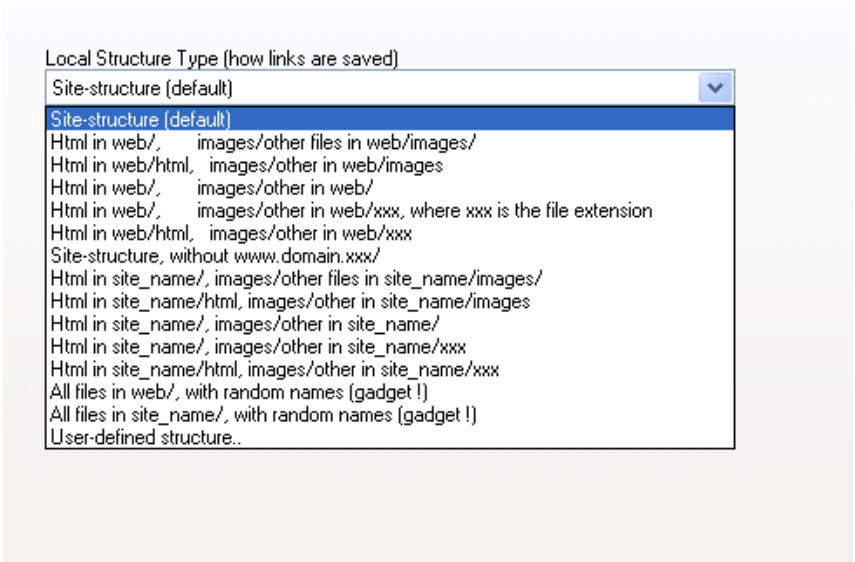
รูป 9.19 การคฃตั้งค่าการเชื่อมต่อให้โหลดข้อมูลอย่างไร

Build ใช้กำหนดการเก็บข้อมูลที่โหลดเข้ามา โดยปกติจะเลือกเก็บแบบที่เก็บอยู่ในเว็บไซ้ต์นั้น (Site-structure) คือมีโครงสร้างโฟลเด้อร์ต่างๆเหมือนกับในเว็บไซ้ต์ หรือจะให้แยกเก็บตามประเภทของไฟล์ก็ได้ซึ่งจะมีตัวเลือกแบบต่าง

ดังในรูป No error pages ใช้บอกไม่ให้สร้างเพ่จที่บอกว่า link นั้น error ไม่สามารถดึงเพ่จตามที่ link ไป No external pages ใช้ทำให้ขึ้นคำเตือน "Warning, you need to be online to go to this link.." สำหรับ link ที่ต่อไปภายนอกซึ่งไม่ได้โหลดเข้ามาในเครื่อง อีกตัวเลือกที่ใช้บ่อย คือ Do not purge old files ซึ่งใช้บังคับไม่ให้ลบไฟล์เก่าซึ่งเมื่อโหลดหรือปรับปรุงข้อมูลใหม่แล้วไม่มีอยู่ในเว็บไซ้ต์นั้น ในกรณีที่เราโหลดข้อมูลมาสะสมทีละส่วน เช่น โหลดวารสารทีละเดือนมาเก็บไว้ หากไม่เลือกตัวเลือกนี้ เวลาโหลดข้อมูลเดือนใหม่ข้อมูลเดือนเก่าจะถูกลบทิ้งไปด้วย

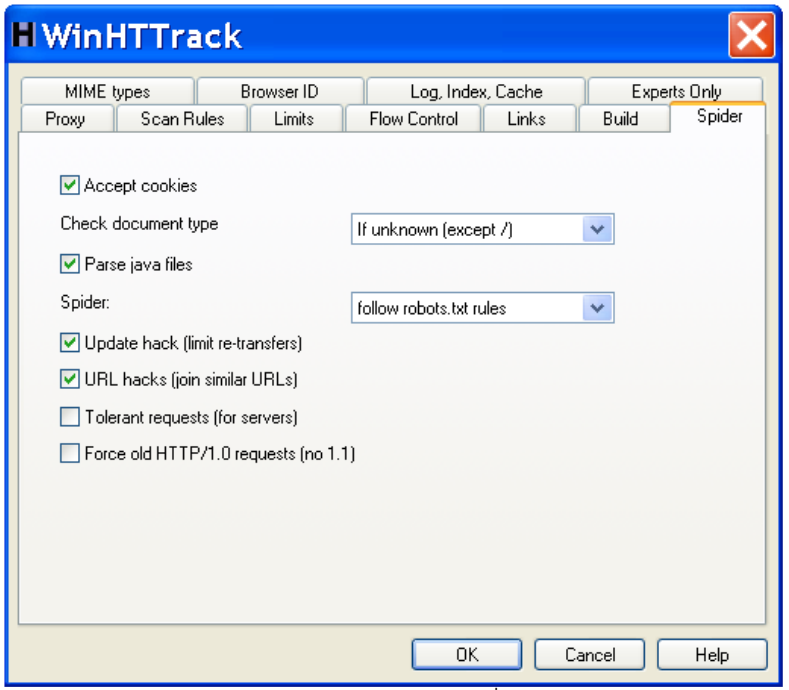


รูป 9.20 การกำหนดวิธีการเก็บข้อมูล



รูป 9.21 กำหนดการเก็บไฟล์ตามแบบโครงสร้างต่างๆ ที่กำหนด

Spider ใช้กำหนดลักษณะการเคลื่อนเข้าหาข้อมูล ปกติไม่จำเป็นต้องเปลี่ยนแปลงหรือตั้งค่าใดใดอีก และกำหนดให้เคารพกฎที่เครื่องแม่ข่ายตั้งไว้ (follow robots.txt rules) แต่ในบางกรณีที่ไม่สามารถโหลดข้อมูลมาได้ก็อาจจำเป็นแก้เป็น no robots.txt rules



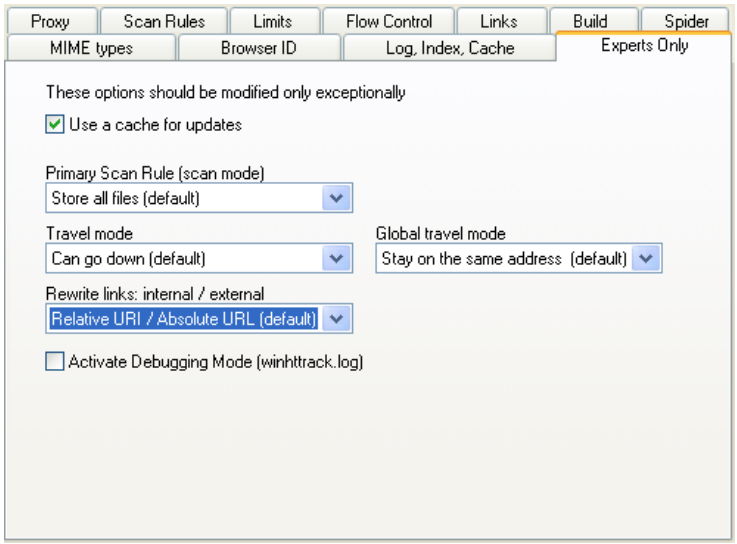
รูป 9.22 การกำหนดลักษณะการเคลื่อนเข้าหาข้อมูล

MIME Types ตัวเลือกนี้เป็นประโยชน์สำหรับเว็บไซ้ต์ที่ไม่ได้สร้างเพ่จเป็นสกุล html แต่เป็น .asp .php ถ้าเรารู้ว่าแท้จริงไฟล์เหล่านี้คือ html เองนั่นเอง ก็ให้ระบุใน MIME type นี้ดังในตัวอย่างข้างล่าง หากไม่ระบุ โปรแกรมจะพยายามทดสอบ link ที่ไปยังไฟล์ .asp .php เหล่านั้นทำให้เสียเวลามากขึ้น

Browser ID ตัวเลือกที่ใช้แสดงตัวกับเครื่องแม่ข่ายว่าเครื่องเรานั้นใช้โปรแกรมใดเชื่อมต่อ และให้สามารถใส่ comment ท้ายไฟล์ที่โหลดมาได้ว่าโหลดมาเมื่อไร ถ้าต้องการประหยัดที่และไม่สนใจข้อมูลนี้ก็เลือกเป็น None ได้

Log, Index, Cache ใช้เก็บข้อมูลการเข้าไปโหลดข้อมูล สร้างไฟล์หน้าแรกสำหรับข้อมูลทั้งหมดที่โหลดมาคือ index.html ในโฟลเด้อร์แรกสุด

Expert Only กำหนดลักษณะการเข้าถึงข้อมูล โดยปกติเราจะโหลดข้อมูลทุกอย่างที่อยู่ในโฟลเด้อร์และที่อยู่ต่ำลงไปทั้งหมด (Can go down + Stay on the same address) แต่บางครั้งที่ link ในหน้าที่ต้องการ link ไปยังเว็บไซ้ต์อื่นๆ เราก็อาจต้องเปลี่ยนตัวเลือกตรงนี้ให้เหมาะสม เช่น “Can both go up and down + Go every where on the web” ซึ่งถ้าเลือกแบบนี้ก็ต้องไม่ลืมจำกัดจำนวนความลึกของการท่องไปในอินเทอร์เน็ตด้วย (Limit – minimum depth)



รูป 9.23 การกำหนดลักษณะการเข้าถึงข้อมูลเว็บไซ้ต์ต่างๆ

6.2 ข้อแนะนำในการใช้โปรแกรม HTTrack

1. โปรแกรมนี้ควรใช้เพื่อการโหลดข้อมูลมาใช้เป็นการส่วนตัวหรือเพื่อการศึกษาเท่านั้น ไม่ควรนำข้อมูลผู้อื่นมากระทำการใดใดอันจะเป็นการละเมิดลิขสิทธิ์แก่เจ้าของเดิมได้

2. ก่อนใช้โปรแกรม ควรท่องเว็บที่ต้องการก่อน เพื่อสำรวจว่ามีข้อมูลที่ตนเองต้องการอยู่ในที่ใดบ้าง link ต่างๆที่ต้องการด้วยนั้นอยู่ภายในเว็บไซ้ต์นั้นและอยู่ในโฟลเด้อร์เดียวกันหรือต่ำลงไปหรือไม่ หากเป็นเช่นนั้นก็สามารถโหลดโดยไม่ต้องปรับเปลี่ยนตัวเลือกใดใดก็ได้ แต่หากมีการเก็บข้อมูลจากหน้าที่ต้องการลงไปอย่างไม่เป็นระบบ เช่น เชื่อมไปยังเว็บไซ้ต์อื่น เชื่อมไปยังไฟล์ในโฟลเด้อร์ที่สูงกว่า ก็ต้องมองรูปแบบและกำหนดตัวเลือกใหม่ให้เหมาะสมกับแต่ละงาน

3. ในกรณีที่ข้อมูลที่โหลดมีมากกว่าที่คาดและต้องการยกเลิก ให้กดปุ่ม Cancel และกดซ้ำอีกครั้งจนโปรแกรมหยุดการโหลดให้ ในกรณีที่ต้องการโหลดข้อมูลต่อในภายหลังให้เลือก Continue interrupted download แต่หากเป็นการโหลดข้อมูลเฉพาะที่ปรับปรุงใหม่ให้เลือก Updating existing download

4. ในกรณีที่เป็นการโหลดข้อมูลทีละส่วนมาสะสมในชื่อโครงการเดียวกัน เช่น โหลดหนังสือพิมพ์ในแต่ละวัน ในแต่ละครั้งก็จะได้ข้อมูลโฟลเด้อร์ของหนังสือพิมพ์วันนั้น วันต่อมาก็โหลดของอีกวันหนึ่ง ต่อไปเรื่อยๆ กรณีเช่นนี้ต้องไม่ลืมเปลี่ยนตัวเลือก Build ให้เป็น Do not purge old files ด้วย ไม่เช่นนั้นเมื่อโหลดข้อมูลครั้งใหม่ ข้อมูลของเก่าจะถูกลบทิ้งไปด้วย

5. ข้อมูลที่โหลดมาเป็นไฟล์ html แล้ว ถ้าต้องการนำมาใส่ในคลังข้อมูลภาษาที่ต้องการก็สามารถแปลงเป็นไฟล์ .txt ได้โดยการใช้โปรแกรม DeHTML ดังที่กล่าวมาแล้ว

7. พระราชบัญญัติลิขสิทธิ์พ.ศ. ๒๕๓๗

การคัดลอกข้อมูลภาษาจากแหล่งต่างๆมาใช้นั้น ควรจะต้องพิจารณาถึงปัญหาทางกฎหมายในเรื่องของลิขสิทธิ์ด้วย ในส่วนภาษาไทยกฎหมายที่เกี่ยวข้องโดยตรง คือ พระราชบัญญัติลิขสิทธิ์พ.ศ.๒๕๓๗ ซึ่งมีสาระสำคัญคือห้ามการ ทำซ้ำ หรือ ดัดแปลง งานอันมีลิขสิทธิ์และเผยแพร่แก่สาธารณะ

"ทำซ้ำ" หมายความรวมถึง คัดลอกไม่ว่าด้วยวิธีใด ๆ เลียนแบบ ทำสำเนา ทำ แม่พิมพ์ บันทึกเสียง บันทึกภาพ หรือบันทึกเสียงและภาพ จากต้นฉบับ จากสำเนาหรือจากการโฆษณา ในส่วนอันเป็นสาระสำคัญ ทั้งนี้ ไม่ว่าทั้งหมดหรือบางส่วน สำหรับในส่วนที่เกี่ยวกับโปรแกรมคอมพิวเตอร์ให้หมายความถึงคัดลอกหรือทำสำเนาโปรแกรมคอมพิวเตอร์จากสื่อบันทึกใด ไม่ว่าด้วยวิธีใด ๆ ในส่วน อันเป็นสาระสำคัญ โดยไม่มีลักษณะเป็นการจัดทำงานขึ้นใหม่ ทั้งนี้ ไม่ว่าทั้งหมดหรือบางส่วน

"ดัดแปลง" หมายความว่า ทำซ้ำโดยเปลี่ยนรูปใหม่ ปรับปรุงแก้ไขเพิ่มเติม หรือ จำลองงานต้นฉบับในส่วนอันเป็นสาระสำคัญโดยไม่มีลักษณะเป็นการจัดทำงานขึ้นใหม่ ทั้งนี้ ไม่ว่า ทั้งหมดหรือบางส่วน

"เผยแพร่ต่อสาธารณชน" หมายความว่า ทำให้ปรากฏต่อสาธารณชน โดย การแสดง การบรรยาย การสวด การบรรเลงการทำให้ปรากฏด้วยเสียงและหรือภาพ การก่อสร้าง การจำหน่าย หรือโดยวิธีอื่นใดซึ่งงานที่ได้จัดทำขึ้น

อย่างไรก็ตาม พ.ร.บ.นี้ได้กำหนดข้อยกเว้นไว้ว่า งานเขียนบางอย่างไม่มีลิขสิทธิ์ และงานเก่าๆ ที่เจ้าของลิขสิทธิ์ได้เสียชีวิตแล้ว 50 ปีไปแล้ว ก็จะสามารถนำมาใช้ได้ หรือหากเป็นการใช้ในบางลักษณะก็มีข้ออนุโลมให้ได้ เช่น การใช้เพื่อการศึกษาวิจัยงานนั้นโดยไม่แสวงหากำไร หรือเพื่อประโยชน์ในการสอน ดังมีรายละเอียดดังนี้

มาตรา ๗ สิ่งต่อไปนี้ไม่ถือว่าเป็นงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้

(๑) ข่าวประจำวัน และข้อเท็จจริงต่าง ๆ ที่มีลักษณะเป็นเพียงข่าวสารอันมิใช่งานในแผนกวรรณคดี แผนกวิทยาศาสตร์ หรือแผนกศิลปะ

(๒) รัฐธรรมนูญ และกฎหมาย

(๓) ระเบียบ ข้อบังคับ ประกาศ คำสั่ง คำชี้แจง และหนังสือตอบโต้ของกระทรวง ทบวง กรม หรือหน่วยงานอื่นใดของรัฐหรือของท้องถิ่น

(๔) คำพิพากษา คำสั่ง คำวินิจฉัย และรายงานของทางราชการ

(๕) คำแปลและการรวบรวมสิ่งต่าง ๆ ตาม (๑) ถึง (๔) ที่กระทรวง ทบวง กรม หรือหน่วยงานอื่นใดของรัฐหรือของท้องถิ่นจัดทำขึ้น

มาตรา ๑๙ ภายใต้บังคับมาตรา ๒๑ และมาตรา ๒๒ ลิขสิทธิ์ตามพระราชบัญญัตินี้ให้มีอยู่ตลอดอายุของผู้สร้างสรรค์และมีอยู่ต่อไปอีกเป็นเวลาห้าสิบปี นับแต่ผู้สร้างสรรค์ถึงแก่ความตาย

ในกรณีที่มีผู้สร้างสรรค์ร่วมลิขสิทธิ์ในงานดังกล่าวให้มีอยู่ตลอดอายุของผู้สร้างสรรค์ร่วม และมีอยู่ต่อไปอีกเป็นเวลาห้าสิบปีนับแต่ผู้สร้างสรรค์ร่วมคนสุดท้ายถึงแก่ความตาย

ถ้าผู้สร้างสรรค์หรือผู้สร้างสรรค์ร่วมทุกคนถึงแก่ความตายก่อนที่ได้มีการโฆษณางานนั้นให้ลิขสิทธิ์ดังกล่าวมีอายุห้าสิบปีนับแต่ได้มีการโฆษณาเป็นครั้งแรก

ในกรณีที่ผู้สร้างสรรค์เป็นนิติบุคคล ให้ลิขสิทธิ์มีอายุห้าสิบปีนับแต่ผู้สร้างสรรค์ได้สร้างสรรค์ขึ้น แต่ถ้าได้มีการโฆษณางานนั้นในระหว่างระยะเวลาดังกล่าว ให้ลิขสิทธิ์มีอายุห้าสิบปีนับแต่ได้มีการโฆษณาเป็นครั้งแรก

มาตรา ๒๗ การกระทำอย่างใดอย่างหนึ่งแก่งานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ โดยไม่ ได้รับอนุญาตตามมาตรา ๑๕ (๕)ให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้กระทำดังต่อไปนี้

(๑) ทำซ้ำหรือดัดแปลง

(๒) เผยแพร่ต่อสาธารณชน

มาตรา ๓๒ การกระทำแก่งานอันมีลิขสิทธิ์ของบุคคลอื่นตามพระราชบัญญัตินี้ หากไม่ขัดต่อการแสวงหาประโยชน์จากงานอันมีลิขสิทธิ์ตามปกติของเจ้าของลิขสิทธิ์ และไม่กระทบกระเทือนถึงสิทธิอันชอบด้วยกฎหมายของเจ้าของลิขสิทธิ์เกินสมควร มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ภายใต้บังคับบทบัญญัติในวรรคหนึ่ง การกระทำอย่างหนึ่งอย่างใดแก่งานอันมีลิขสิทธิ์ตามวรรคหนึ่ง มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้กระทำดังต่อไปนี้

(๑) วิจัยหรือศึกษางานนั้น อันมิใช่การกระทำเพื่อหากำไร

(๒) ใช้เพื่อประโยชน์ของตนเอง หรือเพื่อประโยชน์ของตนเองและบุคคลอื่นในครอบครัวหรือญาติสนิท

(๓) ติชม วิจารณ์ หรือแนะนำผลงานโดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น

(๔) เสนอรายงานข่าวทางสื่อสารมวลชนโดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น

(๕) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ เพื่อประโยชน์ในการพิจารณาของศาลหรือเจ้าพนักงานซึ่งมีอำนาจตามกฎหมาย หรือในการรายงานผลการพิจารณาดังกล่าว

(๖) ทำซ้ำ ดัดแปลง นำออกแสดง หรือทำให้ปรากฏ โดยผู้สอนเพื่อประโยชน์ ในการสอนของตน อันมิใช่การกระทำเพื่อหากำไร

(๗) ทำซ้ำ ดัดแปลงบางส่วนของงาน หรือตัดทอน หรือทำบทสรุปโดยผู้สอน หรือสถาบันศึกษาเพื่อแจกจ่ายหรือจำหน่ายแก่ผู้เรียนในชั้นเรียนหรือในสถาบันศึกษา ทั้งนี้ ต้องไม่เป็นการกระทำเพื่อหากำไร

(๘) นำงานนั้นมาใช้เป็นส่วนหนึ่งในการถามและตอบในการสอบ

มาตรา ๓๓ การกล่าว คัด ลอก เลียน หรืออ้างอิงงานบางตอนตามสมควรจากงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ โดยมีการรับรู้ถึงความเป็นเจ้าของลิขสิทธิ์ในงานนั้น มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์ ถ้าได้ปฏิบัติตามมาตรา ๓๒ วรรคหนึ่ง

มาตรา ๓๔ การทำซ้ำโดยบรรณารักษ์ของห้องสมุดซึ่งงานอันมีลิขสิทธิ์ตามพระราชบัญญัตินี้ มิให้ถือว่าเป็นการละเมิดลิขสิทธิ์หากการทำซ้ำนั้นมิได้มีวัตถุประสงค์เพื่อหากำไร และได้ปฏิบัติตามมาตรา ๓๒ วรรคหนึ่ง ในกรณีดังต่อไปนี้

(๑) การทำซ้ำเพื่อใช้ในห้องสมุดหรือให้แก่ห้องสมุดอื่น

(๒) การทำซ้ำงานบางตอนตามสมควรให้แก่บุคคลอื่น เพื่อประโยชน์ในการวิจัยหรือการศึกษา

บทที่ 10 การกำกับข้อมูล

1. เหตุผลที่มีการกำกับข้อมูล

โดยทั่วไปเอกสารอิเล็กทรอนิกส์ที่ใช้ในงานคลังข้อมูลภาษาจะอยู่ในรูปของข้อความล้วน (plain text) หรือเป็นแฟ้มข้อมูลแบบ ASCII คือเป็นแฟ้มข้อมูลที่ตัวอักษรแต่ละตัวจะเก็บไว้ในคอมพิวเตอร์ด้วยรหัสที่กำหนดตามตารางรหัส ASCII ซึ่งมี 256 รหัสตั้งแต่รหัสที่ 0 ถึง รหัสที่ 255 ข้อมูลที่อยู่ในรูปของข้อความล้วนนี้สามารถที่จะแลกเปลี่ยนข้อมูลกันได้ถึงแม้จะใช้เครื่องคอมพิวเตอร์ต่างชนิดกัน แต่มีข้อเสียคือเอกสารที่เป็น ASCII ไม่สามารถให้ข้อมูลอื่นๆนอกเหนือจากข้อความในเอกสาร เช่น การจัดรูปแบบตัวอักษร ขนาด หรือลักษณะต่างๆของการแสดงหน้าจอหรือพิมพ์ และไม่สามารถแสดงข้อความภาษาอื่นๆที่ไม่ครอบคลุมในรหัส ASCII นี้ เพราะเป็นเพียงข้อมูลดิบที่บอกเพียงว่าในเอกสารนั้นประกอบด้วยตัวอักษรอะไรบ้าง หากต้องการบอกถึงรูปแบบการแสดงผลหรือพิมพ์ข้อความในลักษณะต่างๆ จึงจำเป็นต้องมีการกำกับข้อมูลบางอย่างเพิ่มเข้าไป ดังเห็นได้จากแฟ้มข้อมูลที่พิมพ์ด้วยโปรแกรม word processor ทั้งหลาย จะมีการกำกับรหัสพิเศษเพื่อใช้ในการจัดรูปแบบ (format) แทรกอยู่ด้วย เช่น กำกับว่าส่วนใดเป็นตัวหนา ตัวเอียง อยู่กึ่งกลาง ขนาดเท่าใด เป็นต้น แต่ข้อเสียของวิธีนี้คือ แฟ้มข้อมูลที่สร้างขึ้นจากแต่ละโปรแกรม เช่น MS Word, Ami Pro, Word Perfect เป็นต้น ต่างก็ใช้วิธีกำกับข้อมูลในรูปแบบเฉพาะของตัวเอง ทำให้มีปัญหาสำหรับการแลกเปลี่ยนข้อมูลกันหากไม่ได้ใช้โปรแกรม Word Processor ชนิดเดียวกัน ในทางปฏิบัติจะต้องมีโปรแกรมสำหรับแปลงข้อมูลจากรูปแบบ

หนึ่งไปสู่อีกรูปแบบหนึ่ง ซึ่งโปรแกรม Word Processor ต่างๆ แม้จะยอมให้มีการแปลงข้อมูลจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่ง แต่ก็ไม่แน่นอนเสมอไปว่าเมื่อแปลงข้อมูลแล้วจะสามารถแสดงผลข้อมูลทุกอย่างได้เหมือนต้นฉบับเดิม ในภายหลัง การกำกับข้อมูลจึงเปลี่ยนไป แทนที่จะกำกับลักษณะรูปแบบข้อมูลที่ต้องการก็เปลี่ยนมาเป็นกำกับอธิบายว่าแต่ละส่วนของข้อมูลคืออะไร เช่น เป็นชื่อเรื่อง (title) เป็นหัวเรื่อง (heading) เป็นเนื้อความ (body) เป็นต้น ซึ่งทำให้การแลกเปลี่ยนข้อมูลเป็นไปได้สะดวกมากขึ้น และยังช่วยให้สามารถใช้ประโยชน์จากเอกสารอิเล็กทรอนิกส์ได้ดีมากขึ้น

ในทางภาษาศาสตร์เอง การกำกับข้อมูลก็มีความสำคัญ ทั้งนี้เพื่อใช้ประโยชน์ในการวิเคราะห์ทางภาษาศาสตร์ เพราะคลังข้อมูลภาษาที่มีเพียงข้อความล้วนอย่างเดียวไม่สามารถใช้ประโยชน์ในการวิเคราะห์ภาษาได้อย่างเต็มที่ การใช้คลังข้อมูลที่เป็นข้อความล้วนทำให้สามารถศึกษาได้เฉพาะเรื่องทางภาษาศาสตร์ที่อาศัยแต่รูปคำเป็นหลัก เช่น ศึกษารูปแบบการใช้ของคำที่ต้องการ แต่ถ้าผู้วิจัยต้องการศึกษาเรื่องในระดับที่สูงขึ้นไป เช่น ในระดับวากยสัมพันธ์ คลังข้อมูลภาษาก็จะต้องมีข้อมูลพื้นฐานทางวากยสัมพันธ์ เช่น หมวดคำอยู่ด้วย จึงจำเป็นต้องกำกับข้อมูลทางภาษาศาสตร์บางอย่างเข้าไปในคลังข้อมูลภาษาด้วย การกำกับข้อมูลอื่นๆทางภาษาศาสตร์ที่ไม่ใช่ตัวข้อความเข้าไปในเอกสารนี้เรียกว่า การกำกับข้อมูลทางภาษาศาสตร์ หรือการทำ linguistic annotation

2. การกำกับข้อมูลคืออะไร

การกำกับข้อมูลคือการเพิ่มเติมข้อมูลอื่นๆ นอกเหนือจากตัวข้อความของเอกสารนั้น ข้อมูลส่วนที่เพิ่มเติมเข้าไปนี้เรียกว่า Markup แต่เดิม markup หมายถึงเครื่องหมายหรือสัญลักษณ์ต่างๆที่เขียนเพิ่มเติมในเอกสารเพื่อบอกว่าจะ

ให้จัดพิมพ์เอกสารนั้นๆออกมาอย่างไรเมื่อส่งเรียงพิมพ์ เช่น ให้จัดว่าส่วนไหนจะพิมพ์ตัวหนา ตัวเอียง ตัวใหญ่ markup ประเภทนี้เราเรียกว่า procedural markup เพราะเป็นการกำหนดว่าจะให้ทำอะไรกับข้อมูล ณ ตำแหน่งนั้น ปัจจุบัน markup แบบนี้ยังคงใช้ในโปรแกรม word processor ต่างๆ เพียงแต่ว่าแต่ละโปรแกรมก็จะกำหนดใช้วิธีการ markup ที่แตกต่างกัน การย้ายข้อมูลจากโปรแกรมหนึ่งไปยังอีกโปรแกรมหนึ่งจึงจำเป็นต้องอาศัยโปรแกรมเฉพาะที่ทำการแปลงรูปแบบให้ถูกต้อง

Markup อีกประเภทเรียกว่า descriptive markup เป็นการ markup เพื่อบอกโครงสร้างของเอกสารนั้นๆ ไม่ใช่การ markup เพื่อบอกรูปแบบที่จะปรากฏแสดงหน้าจอหรือพิมพ์เหมือนอย่างใน procedural markup descriptive markup จะบอกโครงสร้างของเอกสารเช่น chapter, section, table of content, เป็นต้น ซึ่งจะทำให้เอกสารนั้นเป็นอิสระจากสื่อที่จะปรากฏ เมื่อนำเอกสารนั้นไปพิมพ์เป็นหนังสือ พิมพ์ลงในซีดีรอมหรือ เผยแพร่ในอินเทอร์เน็ตก็สามารถทำได้ทันทีโดยไม่ต้องแปลงเอกสารนั้น (การจัดรูปแบบ (format) ตามที่ต้องการสำหรับสื่อแต่ละประเภท จะถูกกำหนดจากโปรแกรมต่างๆ ที่ใช้งานข้อมูลนั้น รูปแบบไม่ได้กำหนดไว้ในตัวเอกสาร) ด้วยวิธีนี้ เอกสารก็จะมีลักษณะที่เป็นกลางมากขึ้นทำให้มีการแลกเปลี่ยนข้อมูลกันระหว่างคอมพิวเตอร์ได้ง่ายขึ้น ถ้าไม่ใช้ descriptive markup แต่ไปใช้ procedural markup การแปลงเอกสารจากรูปแบบหนึ่งไปอีกรูปแบบหนึ่งก็จะเสียเวลาและยุ่งยากมากขึ้น

3. หน้าที่และบทบาทของ TEI

TEI ย่อมาจาก Text Encoding Initiative เป็นโครงการวิจัยนานาชาติ (international research project) ที่ตั้งขึ้นมาเพื่อพัฒนาและวางแนวทางสำหรับการกำกับข้อมูลเพื่อให้การแลกเปลี่ยนเอกสารอิเล็กทรอนิกส์เป็นไปได้อย่างอิสระ

โดยไม่มีข้อจำกัดด้านฮาร์ดแวร์คอมพิวเตอร์หรือซอฟแวร์ที่ใช้ หน่วยงานที่ให้การสนับสนุน TEI คือ ACH (The Association for Computers and the Humanities), ACL (The Association for Computational Linguistics) และ ALLC (The Association for Literary and Linguistics Computing)

โครงการของ TEI เริ่มในปีค.ศ.1987 ในการประชุมทางวิชาการที่ Vassar College นักวิชาการ 15 คนจากอเมริกาเหนือและยุโรปได้รับเลือกให้เป็นกรรมการบริหาร (advisory board) หนังสือคู่มือแนวทางการกำกับข้อมูลภาษาของ TEI ฉบับสมบูรณ์ได้รับการเผยแพร่เป็นทางการเมื่อเดือนพฤษภาคม ปีค.ศ.1994 เรียกว่า TEI P3 (ก่อนหน้านั้นมีการเผยแพร่ฉบับร่าง TEI P1 ในปีค.ศ. 1990 และ TEI P2 ในปีค.ศ.1992)

เมื่อเริ่มต้น TEI ได้เสนอให้ใช้ SGML เป็นมาตรฐานในการกำกับข้อมูลเอกสารอิเล็กทรอนิกส์ คณะกรรมการของ TEI ได้ร่วมประชุมและเสนอรายการแท็กพื้นฐานและแฟ้มข้อมูล DTD (document type definition) สำหรับใช้กับเอกสารแบบต่างๆ แท็กที่ใช้ใน SGML จะอยู่ภายในเครื่องหมาย <...> โดยที่มี <...> เป็นตัวบอกจุดเริ่มต้น และ </...> เป็นตัวบอกจุดสิ้นสุดของข้อมูลในแท็กนั้น ข้อมูลต่างๆ ของ TEI ได้เผยแพร่แก่สาธารณะทางอินเทอร์เน็ตที่ http://www.tei-c.org/index.xml ต่อมาในปีค.ศ.1998 เมื่อ World Wide Web Consortium ได้แนะนำมาตรฐาน XML .ให้เป็นมาตรฐานอนาคตของเอกสารเว็บ ทางกรรมการบริหาร TEI จึงได้มีการปรับการทำงานใหม่จัดตั้งเป็น TEI consortium และได้ปรับมาตราฐานการกำกับข้อมูลใหม่ให้ใช้ XML ซึ่งเป็นมาตรฐานที่คล้ายกับ SGML แต่ชัดเจนและง่ายต่อการประมวลผลมากกว่า หนังสือคู่มือแนวทางการกำกับข้อมูลภาษาของ TEI แบบที่ใช้ XML คือ TEI P4 ออกเผยแพร่ในปีค.ศ.2002 ในปัจจุบัน หนังสือคู่มือฉบับล่าสุดคือ TEI P5 ซึ่ง

ออกเป็นทางการในเดือนพฤศจิกายน ค.ศ.2007 ในคู่มือฉบับนี้ การกำกับข้อมูลภาษาจะเปลี่ยนไปจากเดิมมาก เพราะมีการแยกระหว่างตัวข้อมูลภาษากับตัวข้อมูลการกำกับออกจากกัน อย่างไรก็ตาม การแปลงข้อมูลที่กำกับตามแบบ TEI P4 จะสามารถถ่ายโอนไปเป็นแบบ TEI P5 ได้ด้วยการใช้โปรแกรมคอมพิวเตอร์ที่ TEI จัดทำขึ้น ในบทนี้จะกล่าวถึงการกำกับข้อมูลตามแบบ TEI P4 เป็นหลัก

4. SGML และ XML

SGML ย่อมาจาก Standard Generalized Markup Language ถูกกำหนดเป็นมาตรฐาน ISO ในปีค.ศ.1986 เพื่อใช้เป็นมาตรฐานสำหรับการกำกับข้อมูล บุคคลที่มีความสำคัญต่อการพัฒนาของ SGML คือ Charles Goldfarb ในปีค.ศ.1969 ในขณะที่เขาทำงานที่บริษัท IBM เขาได้พัฒนา GML (Generalized Markup Language) ร่วมกับ Edward Mosher และ Raymond Lorie เพื่อใช้ในการจัดรูปแบบเอกสาร (text formatting) การแก้ไขเอกสาร (editting) และการสืบค้นข้อมูล (information retrieval) ต่อมา ในปี ค.ศ. 1978 American National Standards Institute (ANSI) ได้ตั้งคณะกรรมการเพื่อวางมาตรฐานภาษาสำหรับใช้กำกับข้อมูล Goldfarb ได้รับเชิญให้เป็นหนึ่งในคณะกรรมการนี้ และเขาก็ทำงานให้กับ ISO (International Organization for Standardization) ด้วย เขาจึงได้นำแนวคิดจาก GML มาพัฒนาต่อเป็น SGML โครงร่างฉบับแรกของ SGML ซึ่งได้รับการเผยแพร่เป็นทางการในปี ค.ศ. 1985

SGML เป็นข้อกำหนดที่ใช้กำหนดภาษา (meta-language) กล่าวคือ SGML เป็นตัวกำหนดว่าภาษาที่ใช้กำกับข้อมูล (markup language) ควรจะมีคุณสมบัติอย่างไรบ้าง SGML ช่วยให้เราสามารถกำหนดโครงสร้างในระดับต่างๆ (hierarchical structure) ของเอกสารนั้นๆได้ว่ามีองค์ประกอบอะไรบ้าง SGML เป็น descriptive markup ดังนั้น เอกสารที่ถูกกำกับตามมาตรฐาน SGML จะมี

ลักษณะเป็นกลางและสามารถนำไปใช้กับโปรแกรมใดก็ได้ที่รู้จักและสามารถประมวลผลแฟ้มข้อมูลที่เป็นไปตามมาตรฐาน SGML

SGML ถูกออกแบบมาโดยมีเป้าหมายให้เป็นมาตรฐานที่ไม่ขึ้นกับคอมพิวเตอร์ชนิดใดชนิดหนึ่ง ไม่ขึ้นกับอุปกรณ์ที่ใช้เก็บข้อมูล ไม่ขึ้นกับกฎเกณฑ์ใดโดยเฉพาะ และไม่ขึ้นกับงานใดงานหนึ่งโดยเฉพาะ ดังนั้น เอกสารที่ถูกกำกับตามมาตรฐานของ SGML จะสามารถนำไปใช้กับเครื่องคอมพิวเตอร์ประเภทใดก็ได้ จะผลิตออกมาในสื่อแบบใดก็ได้ จะนำชุดกฎใดมาประมวลผลข้อมูลก็ได้ ข้อมูลต่างๆไม่ว่าจะเป็น ตาราง สูตรทางคณิตศาสตร์ สูตรทางเคมี ตัวอย่าง ก็สามารถแสดงและจัดรูปแบบตามที่ต้องการได้ในภายหลังตามที่แต่ละโปรแกรมต้องการจะให้ข้อมูลนั้นแสดงออกมาอย่างไร

อย่างไรก็ตาม มาตรฐาน SGML ไม่เป็นที่นิยมใช้กันอย่างแพร่หลายเท่าที่ควร สาเหตุส่วนหนึ่งมาจากความซับซ้อนของ SGML เองที่ทำให้ใช้งานยาก เห็นได้จากหนังสือคู่มืออ้างอิงของ SGML ที่มีขนาดกว่า 500 หน้า ในปีค.ศ.1996 World Wide Web consortium (W3C) ได้สนับสนุนคณะที่ทำงาน SGML เดิม (SGML Editorial Review Board) ให้ทำการปรับมาตรฐานให้มีความเรียบง่ายมากขึ้น จนได้มาเป็นมาตรฐาน XML และได้นำเสนอแก่ W3C ในปีค.ศ.1998 XML จึงเป็นเหมือนเซ็ทย่อยของ SGML

SGML/XML มองเอกสารเสมือนเป็นวัตถุ (object) อย่างหนึ่ง เอกสารแต่ละประเภท (type) ก็จะมีคุณสมบัติและโครงสร้างที่ต่างกัน เช่น มี section, chapter, paragraph, sentence เอกสารจะประกอบด้วยหน่วย (element) ต่างๆ หน่วยเหล่านี้จะต้องถูกกำกับไว้อย่างชัดเจนในเอกสารนั้น ซึ่งโดยทั่วไปก็จะกำกับด้วยแท็กเปิดและปิดท้ายด้วยแท็กปิดเหมือนกับการใช้เครื่องหมายคำพูด

(quotation) เปิดและปิดข้อความ นอกจากนี้ หน่วยบางหน่วยอาจซ้อนอยู่ภายใต้ (embed) หน่วยอื่นก็ได้ ดังตัวอย่างนี้

```
<anthology>
          <poem><title>The SICK ROSE</title>
             <stanza>
                  <line>O Rose thou art sick.</line>
                  <line>The invisible worm,</line>
                  <line>That flies in the night</line>
                  <line>In the howling storm:</line>
             </stanza>
             <stanza>
                  <line>Has found out thy bed</line>
                  <line>Of crimson joy:</line>
                  <line>And his dark secret love</line>
                  <line>Does thy life destroy.</line>
             </stanza>
          </poem>
                   <!-- more poems go here    -->

</anthology>
```

รูป 10.1 ตัวอย่างของ XML document (จาก TEI P4)

ในตัวอย่างนี้ มีการกำหนดหน่วยที่ใช้ 5 หน่วย คือ anthology, poem, title, stanza, line ภายใน anthology ประกอบด้วย poem ได้มากกว่าหนึ่ง poem ภายใน poem ประกอบด้วย 2 หน่วย คือ title กับ stanza และภายใน stanza ประกอบด้วย line ได้มากกว่าหนึ่ง line

ในปัจจุบันมีภาษาสำหรับกำกับข้อมูล (markup language) ที่พัฒนาขึ้นมาโดยอาศัยมาตรฐานของ SGML และใช้กันอยู่แพร่หลาย เช่น HTML (Hyper Text Markup Language) ซึ่งเป็นภาษาสำหรับกำกับข้อมูลที่ใช้กับเอกสารที่เป็น hypertext และใช้กันอยู่อย่างแพร่หลายในอินเทอร์เน็ต หรือภาษา CDIF (Corpus Document Interchange Format) ซึ่งเป็นภาษาสำหรับกำกับข้อมูลที่ใช้กับคลังข้อมูล BNC เป็นต้น อย่างไรก็ตาม W3C ได้วางอนาคตไว้ว่า HTML ก็จะเปลี่ยนไปเป็น XML ในที่สุด ในที่นี้จึงจะกล่าวถึงมาตรฐาน XML เป็นหลัก

4.1 องค์ประกอบของเอกสาร SGML/XML

ในการสร้างเอกสารที่เป็น SGML/XML (SGML/XML document) มีองค์ประกอบอยู่สามส่วน คือ ส่วน SGML/XML declaration, Document Type Definition (DTD) และตัว SGML/XML document instance ส่วนที่สองและสามนั้นมีลักษณะร่วมกัน โดยมาตรฐาน XML จะจำกัดกว่า SGML (XML เป็นเซ็ทย่อยของ SGML) ในที่นี้จึงจะอธิบายโดยอ้างอิงเฉพาะสิ่งที่มีใน XML เป็นหลัก

4.1.1 SGML Declaration

ส่วนนี้เป็นส่วนที่กำหนดกลุ่มตัวอักษรที่จะใช้ (character set) ลักษณะไวยากรณ์ (concrete syntax) ของ DTD และคุณลักษณ์ (feature) อื่นๆ ของ SGML ที่ใช้ในระบบนั้น เช่น การกำหนดเรื่องของ minimization, link, concur, และ subdoc รูปที่ 10.2 แสดง SGML Declaration ของเอกสาร SGML แบบพื้นฐาน คือ ในกรณีที่ไม่มีการกำหนด SGML declaration สำหรับระบบ ก็ให้ถือเอาค่าพื้นฐานตามนี้

```
<!SGML "ISO 8879:1986"
       -- Declaration for typical Basic SGML Document --
  CHARSET  BASESET   "ISO 646:1983//CHARSET International
                      Reference Version (IRV)//ESC 2/5 4/0"
           DESCSET   0   9  UNUSED
                     9   2  9
                     11  2  UNUSED
                     13  1  13
                     14  18 UNUSED
                     32  95 32
                     127 1  UNUSED
  CAPACITY PUBLIC    "ISO 8879:1986//CAPACITY Reference//EN"
  SCOPE    DOCUMENT
  SYNTAX   PUBLIC    "ISO 8879:1986//SYNTAX Reference//EN"
  FEATURES MINIMIZE  DATATAG NO  OMITTAG  YES  RANK     NO
SHORTTAG YES    LINK      SIMPLE  NO  IMPLICIT NO   EXPLICIT NO
OTHER     CONCUR  NO  SUBDOC   NO   FORMAL   NO
  APPINFO  NONE
>
```

รูป 10.2 SGML declaration สำหรับเอกสาร SGML ทั่วไป

ส่วน SGML Declaration เริ่มต้นด้วยเครื่องหมาย <! และปิดท้ายด้วยเครื่องหมาย > หลังจากเครื่องหมายเปิดจะตามด้วย SGML “ISO 8879:1986” ซึ่งบอกหมายเลข ISO และปีค.ศ.ที่เสนอมาตรฐาน บรรทัดที่สองเป็นส่วนของการอธิบายเพิ่มเติม (comments) ซึ่งจะเป็นข้อความที่อยู่ระหว่างเครื่องหมาย - ภายใน SGML Declaration ประกอบด้วย 6 ส่วน ได้แก่

- CHARSET ใช้กำหนดชุดอักขระ (character set) ที่ใช้ในเอกสาร SGML ค่าตั้งต้น (default) คือ ชุดอักขระแบบ ISO 646 แต่ก็สามารถเพิ่มเติมชุดอักขระ ISO อื่นๆ ได้ โดยการกำหนดในส่วน DESCSET ของ SYNTAX
- CAPACITY ใช้กำหนดขนาดของ markup character ทั้งหมดว่าไม่เกินค่าที่กำหนด ค่าตั้งต้นเป็น 35,000 ตัวอักษร ข้อกำหนดนี้มีขึ้นเนื่องมาจากข้อจำกัดของคอมพิวเตอร์ในช่วงทศวรรษที่ 1980 ในปัจจุบันส่วนนี้จึงไม่มีความสำคัญนัก
- SCOPE ใช้บอกว่าข้อกำหนด SYNTAX ใช้กับทั้งส่วน document instance และส่วนต้นของเอกสาร (document prolog) ด้วย หรือใช้เฉพาะกับ document instance อย่างเดียว
- SYNTAX เป็นส่วนที่สำคัญที่สุดใน SGML Declaration ใช้กำหนดรายละเอียดของชุดอักขระที่ใช้ โดยกำหนดว่ารหัสอักขระ (character code) ใดบ้างเป็นอักขระควบคุม (control code) รหัสใดบ้างเป็นอักขระปกติ (base character) รหัสใดบ้างใช้ในหน้าที่เฉพาะ (function character) เช่น ใช้เป็น separator ในส่วน declaration เป็นต้น นอกจากนี้ ส่วน SYNTAX นี้ยังกำหนดอักขระที่สามารถใช้ในการตั้งชื่อแท็ก ในการตั้งชื่อ entity และที่ใช้เป็น delimiter เช่น & % < </ เป็นต้น และกำหนดความยาวสูงสุดของสิ่งต่างๆ เช่น ชื่อแท็ก ชื่อ entity เป็นต้น

- FEATURE เป็นส่วนที่กำหนดคุณลักษณ์ของ SGML เช่น ให้มีการละแท็กได้หรือไม่ ให้มีการใช้การอ้างอิงโดยย่อ (short reference) ได้หรือไม่ ให้มีการใช้โครงสร้างซ้อน (concurrent structure) ได้หรือไม่ เป็นต้น
- APPINFO เป็นส่วนที่กำหนดว่าใน SGML application นั้นจำเป็นต้องใช้ application อื่นๆด้วยหรือไม่ เช่น Hytime (Hypermedia/Time-based Structuring Language) หากไม่จำเป็นต้องใช้ application อื่นใดค่าจะเป็น NONE

4.1.2 XML Declaration

XML declaration เป็นส่วนที่กำหนดการประมวลผลข้อมูล จะใส่ไว้บรรทัดแรกสุดของเอกสาร XML ในเวอร์ชั่นปัจจุบัน (1.1) บังคับว่าต้องใส่ไว้ในขณะที่เวอร์ชั่นแรก (1.0) ยอมให้ละได้ XML declaration หรือที่เรียกว่า Processing Instruction มีรูปแบบดังนี้

<?xml version="version_number" encoding="encoding_declaration"
standalone="standalone_status" ?>

โดยที่ version บอกถึงเวอร์ชั่นของ XML ที่ใช้ เช่น 1.1 encoding บอกการเข้ารหัสตัวอักษรว่าใช้มาตรฐานรหัสใด เช่น iso-8859-1 ส่วน standalone มีค่าเป็น yes ถ้าหากว่าข้อมูล DTD ถูกบรรจุไว้ในเอกสาร XML นั้นด้วย

4.1.3 Document instance

คือส่วนที่เป็นตัวเอกสารซึ่งจะถูกกำกับด้วยแท็กต่างๆ แท็กจะอยู่ในรูปของ <Tag_name> และ</Tag_name> เพื่อใช้บอกจุดเริ่มต้นและจุดสิ้นสุดของข้อมูลส่วนนั้น เช่น <title> TEI Tutorial no 2: SGML </title> เป็นข้อมูลส่วนที่ถูกกำกับว่าเป็น title ของเอกสารนั้น ในแท็กต่างๆ อาจจะกำหนดคุณสมบัติ

(attribute) ที่เป็นไปได้ของข้อมูลหน่วยนั้นด้วยก็ได้ รูปแบบของแท็กเปิดจะเป็น <Tag_name Att_set> โดยที่ Att_set เป็นรายการคุณสมบัติของแท็กนั้นๆ คุณสมบัติแต่ละอย่างจะอยู่ในรูปของ Att_name = “Att_value” ส่วนแท็กปิดจะอยู่ในรูป </Tag_name> ชื่อของแท็กและคุณสมบัติของแท็กแต่ละแท็กจะถูกกำหนดไว้ในส่วนที่เป็น DTD DTD จึงเป็นส่วนที่กำหนดลักษณะโครงสร้างของเอกสาร

4.1.4 Document Type Definition (DTD)

หน้าที่ของ DTD ก็คือ กำหนดชื่อแท็กต่างๆ คุณสมบัติต่างๆของแต่ละแท็กและโครงสร้างของแท็กทั้งหมดที่ใช้ เนื่องจาก XML มองเอกสาร (document) ว่าเป็นวัตถุ (object) อย่างหนึ่งซึ่งมีคุณสมบัติประจำตัว ดังนั้นเอกสารประเภทเดียวกันก็ควรจะมีคุณสมบัติและโครงสร้างที่เหมือนกัน หรือกล่าวอีกนัยหนึ่งได้ว่าสามารถกำหนดให้ใช้ DTD เดียวกันได้ ส่วนเอกสารที่ต่างประเภทกัน เช่น บันทึกช่วยจำ รายงานการประชุม เรื่องสั้น ก็จะมีโครงสร้างและองค์ประกอบที่แตกต่างกัน หรือมี DTD ที่แตกต่างกัน คณะกรรมการของ TEI จึงได้กำหนดแนวทางการออกแบบ DTD ของเอกสารแต่ละประเภทเอาไว้แล้ว ผู้ที่จะกำกับข้อมูลภาษาตามมาตรฐาน XML จึงควรศึกษาดู DTD ที่ TEI ได้กำหนดไว้ก่อนเพื่อเป็นแนวทางสำหรับสร้าง DTD สำหรับเอกสารประเภทที่ตนเองต้องการ

ตัวอย่างของสิ่งที่สามารถกำหนดไว้ใน DTD เช่น

- กำหนดชื่อและเนื้อหาของทุกหน่วยในเอกสารประเภทนั้น เช่น เอกสารประเภทบันทึกช่วยจำ จะต้องมีหน่วย To, From, และ Body เป็นต้น
- กำหนดว่าแต่ละหน่วยนั้นปรากฏได้มากน้อยเพียงใด
- กำหนดลำดับในการปรากฏของแต่ละหน่วย

- กำหนดคุณสมบัติของแต่ละแท็กว่าสามารถมีคุณสมบัติอะไรได้บ้าง ค่าตั้งต้น (default) ของแต่ละคุณสมบัติคืออะไร

ตัวอย่างในรูปที่ 10.3 เป็นส่วนหนึ่งของ DTD ที่ใช้สำหรับเอกสารประเภทโคลงกลอน (ซึ่งแสดงไว้ในรูปที่ 10.1)

```
<!ELEMENT anthology      (poem+)>
<!ELEMENT poem           (title?, stanza+)>
<!ELEMENT title          (#PCDATA) >
<!ELEMENT stanza         (line+)   >
<!ELEMENT line           (#PCDATA) >
```

รูป 10.3 ตัวอย่างของ DTD สำหรับเอกสารประเภทโคลงกลอน (จาก TEI P4)

ส่วนแรกที่ตามหลัง ELEMENT จะเป็นชื่อของแท็กที่ใช้ได้สำหรับเอกสารประเภทนั้น ส่วนต่อมาเป็นส่วนเนื้อหาบอกว่าหน่วยนั้นประกอบด้วยหน่วยย่อยอะไรบ้างในลำดับแบบไหน หรือเป็นข้อมูลประเภทไหน และจำนวนครั้งที่สามารถปรากฏได้ เช่น เครื่องหมาย + หลัง poem ที่อยู่ภายในหน่วย anthology หมายความว่า หน่วย anthology จะต้องมีหน่วยของ poem ปรากฏอย่างน้อยหนึ่งหน่วย สำหรับหน่วย poem นั้น ประกอบด้วยหน่วยย่อยสองหน่วยคือ title และ stanza ตามลำดับ โดยที่หน่วยที่เป็น title นั้น อาจไม่ปรากฏก็ได้ (เครื่องหมาย ? ใช้บอกว่าไม่จำเป็นต้องปรากฏหรือเป็น optional) ส่วนหน่วยที่เป็น stanza นั้นจะต้องปรากฏอย่างน้อยหนึ่งครั้ง ส่วนหน่วย title นั้น อาจละแท็กปิดได้ ภายในหน่วยนี้เป็นข้อมูลที่เป็น #PCDATA (parsed character data) หมายความว่าข้อมูลจะเป็นตัวอักษรอะไรก็ได้ที่เครื่องรู้จัก

ใน DTD นั้น เวลาที่ต้องการเขียนคำอธิบายเพิ่มเติมที่ไม่เกี่ยวข้องกับการกำหนดหน่วยต่างๆ หรือที่เรียกว่า comments จะใช้เครื่องหมาย <!- เป็นตัวเปิดข้อความ และปิดข้อความด้วยเครื่องหมาย -> ข้อความที่อยู่ภายในเครื่องหมายนี้ไม่ถือเป็นการกำหนดส่วนใดใดของ DTD เช่น <!-- This is where you can make any comments -->

รายละเอียดของการกำหนดแบบต่างๆที่จะพบใน DTD มีดังต่อไปนี้คือ

4.1.4.1 การกำหนดรายละเอียด (Declaration)

1. DOCTYPE ใช้กำหนดชื่อของ DTD และ declaration อื่นๆ DOCTYPE จะเป็นส่วนแรกสุดใน declaration สามารถเขียนได้สองแบบ ตัวอย่างเช่น

```
<!doctype     memo    [ List_of_declaration     ]>
<!doctype     memo     system    “c:\memo.dtd” >
```

ทั้งสองแบบเป็นตัวอย่างการเขียน DTD ของเอกสารประเภทบันทึกช่วยจำ (memo) ในแบบแรกจะแสดงรายการของการกำหนดรายละเอียดทั้งหมดไว้ภายใน [ ] แต่แบบที่สองจะแยกการกำหนดรายละเอียดทั้งหมดไว้เป็นแฟ้มข้อมูลต่างหากที่ชื่อว่า “memo.dtd”

2. ELEMENT ใช้กำหนดชื่อของแท็ก การละแท็ก และลักษณะเนื้อหาของแท็กนั้นๆ โดยจะเขียนในรูปของ

```
<!ELEMENT  name  content_declaration  >
```

Element จัดเป็นหน่วยพื้นฐานที่เล็กที่สุดของเอกสาร ลำดับและโครงสร้างของแต่ละหน่วยจะถูกกำหนดไว้ใน DTD โดยการใช้ declaration "ELEMENT" นี้ ส่วนคุณสมบัติของแต่ละหน่วยจะถูกกำหนดโดยการใช้ declaration "ATTLIST"

content declaration ใช้กำหนดว่าจะให้หน่วยนั้นๆ ประกอบด้วยหน่วยย่อยอะไรได้บ้างและมีโครงสร้างอย่างไรหรือมีข้อมูลประเภทไหน ตัวอย่าง เช่น

```
<!ELEMENT  Memo - -  ((To & From), Body, Close?)+(Fn) >
```

ใช้เพื่อกำหนดว่าหน่วยที่ชื่อ Memo จะต้องมีทั้งแท็กเปิดและแท็กปิด และภายใน Memo มีหน่วยย่อยได้อีกถึง 5 หน่วยคือ To, From, Body, Close, และ Fn โดย

ที่ To และ From จะต้องปรากฏทั้งสองหน่วยแต่หน่วยไหนจะปรากฏก่อนก็ได้ (เครื่องหมายวงเล็บและ & ใช้บอกว่าต้องมีทั้งสองแต่อันใดมาก่อนหลังก็ได้) จากนั้นจะตามด้วยหน่วย Body และตามด้วยหน่วย Close ตามลำดับ (เครื่องหมาย , ใช้บอกลำดับการปรากฏ) แต่หน่วย Close นี้อาจจะไม่ปรากฏก็ได้ (เครื่องหมาย ? ใช้บอกว่าไม่จำเป็นต้องปรากฏ) ส่วนหน่วย Fn สามารถปรากฏในตำแหน่งใดใดภายใน Memo ก็ได้ (เครื่องหมาย + ใช้แทน inclusion หมายถึงหน่วยนี้สามารถ ปรากฏที่ใดก็ได้ ในขณะที่เครื่องหมาย - จะใช้แทน exclusion หมายถึงหน่วยนี้จะต้องไม่ปรากฏที่ใดเลย)

3. ATTLIST ใช้กำหนดคุณสมบัติของแท็กแต่ละแท็ก จะเขียนในรูปของ

```
<!ATTLIST Element_name Att_name Declare_value Default_value >
```

Element_name เป็นชื่อของแท็กที่ต้องการกำหนดคุณสมบัติ Att_name เป็นชื่อของคุณสมบัติ (attribute) ของแท็กนั้น ค่าของคุณสมบัติจะต้องเป็นไปตามที่กำหนดไว้ใน Declare_value ถ้าหากว่าในข้อมูลในเอกสารนั้นไม่ได้ระบุค่าของ attribute ไว้ก็ให้ใช้ค่าที่กำหนดไว้ใน Default_value ตัวอย่างเช่น

```
<!ATTLIST poem
          id  ID  #IMPLIED
          status (draft | revised | published) "draft">
```

ในตัวอย่างข้างบนนี้ใช้เพื่อแสดงว่าหน่วยที่ชื่อ poem สามารถมีคุณสมบัติได้ 2 อย่าง คือ มี id กับ status ค่าของ id คือตัวเลขหรือชื่อเฉพาะที่กำหนดให้กับ poem หน่วยนั้น (ID หมายถึงข้อมูลที่มีลักษณะเป็น identification) ส่วน #IMPLIED หมายความว่า ค่าของ id นี้ไม่จำเป็นต้องระบุก็ได้ ในกรณีนี้หมายความว่าบาง poem อาจไม่มี id ก็ได้ ส่วนค่าของ status มีได้ 3 อย่างคือ draft, revised และ published หากไม่กำหนดค่าให้ ก็ให้ใช้ค่าที่เป็นอยู่ใน

ปัจจุบัน เช่น ถ้าข้อมูลเป็น <poem status="draft"> ..... </poem> เมื่อโปรแกรมที่จะจัดรูปแบบการพิมพ์พบว่า status ของ poem เป็น draft ก็อาจจะพิมพ์ด้วยตัวอักษรอีกแบบหนึ่ง ต่อมาเมื่อพบข้อมูลใหม่เป็น <poem> ..... </poem> ซึ่งในข้อมูลใหม่นี้ไม่มีการกำหนดคุณสมบัติของ status ไว้ ก็จะถือว่าค่าของคุณสมบัติ status มีค่าตามที่เป็นอยู่คือเป็น draft เหมือนกับ poem ก่อนหน้านี้

นอกจากนี้ หน่วยที่ต่างกันอาจจะใช้ชื่อของคุณสมบัติ (attribute name) เหมือนกันได้ และเรายังสามารถกำหนดคุณสมบัติในหน่วยเพื่อทำการโยงความสัมพันธ์ (cross reference) ได้ ดังในตัวอย่างข้างล่างนี้

สมมติว่า มี document instance ดังนี้

```
<POEM id=Rose>
       Text of poem with identifier 'ROSE'
</POEM>
<POEM id=P40>
       Text of poem with identifier 'P40'
</POEM>

<POEM>
         This poem has no identifier
</POEM>
```

และมี DTD ดังนี้

```
<!ELEMENT poemRef  EMPTY>
<!ATTLIST poemRef   target   IDREF   #REQUIRED >
```

ในตัวอย่างนี้ แท็กที่ชื่อ poemref ถูกกำหนดไว้ใน DTD ไม่ให้มีเนื้อความภายใน (content เป็น EMPTY) และแท็กนี้มีคุณสมบัติชื่อ target โดยที่ค่าที่เป็นได้คือ ค่าที่ถูกใช้เป็น id (IDREF) และจำเป็นต้องมีการระบุค่านี้เสมอ

(เนื่องจากถูกกำหนดให้เป็น #REQUIRED) ซึ่งเราสามารถอ้างถึง poem ที่มี id ว่า Rose ได้โดยใช้แท็ก poemref ดังนี้

```
Blake's poem on the sick rose <poemRef TARGET="Rose"/>
...
```

เมื่อพบข้อมูลเช่นนี้ โปรแกรมจะรู้ว่าข้อความส่วนนี้เชื่อมโยงกับข้อมูลส่วนที่มีเป็น poem และมี id เป็น Rose

4. ENTITY ใช้กำหนด entity ที่จะใช้ในแฟ้มข้อมูล entity เปรียบเหมือนเป็นชื่อที่ใช้เรียกแทนข้อมูลที่กำหนด ปกติจะใช้ entity เพื่อประโยชน์ในการเขียนแบบย่อหรือทำ short hand notation หรือใช้สำหรับอ้างถึงตัวอักขระพิเศษหรือสำหรับอ้างอิงถึงแฟ้มข้อมูลอื่นๆ entity ถูกกำหนดใน DTD โดยการใช้ declaration "ENTITY" เวลาที่อ้างถึง entity ในแฟ้มข้อมูล จะใช้เครื่องหมาย "&" นำหน้า และปิดท้ายด้วย ";" เช่น "&ent1;" เมื่อโปรแกรมที่วิเคราะห์แฟ้มข้อมูลในรูปแบบของ XML พบว่ามีการใช้ entity โปรแกรมจะแทนที่ entity นั้นด้วยค่าเต็มของ entity นั้น การใช้ entity จึงช่วยประหยัดเวลาการพิมพ์ข้อมูลและทำให้สะดวกต่อการดูแลรักษาข้อมูล และช่วยแก้ปัญหาเรื่องตัวอักษรพิเศษได้ ตัวอย่างของ entity เช่น

```
<!ENTITY ChapTwo SYSTEM
"http://www.tei-c.org/P4X/p4chap2.xml">
```

เมื่อกำหนด entity แล้ว แทนที่จะต้องใส่เนื้อความทั้งหมดจากแฟ้มข้อมูล `p4chap2.xml` ลงไป ก็สามารถอ้างถึงเนื้อความในแฟ้มข้อมูลนั้นโดยใช้ entity ChapTwo ได้ ซึ่งเมื่อมีการตีความ The following text has been suppressed: &ChapTwo; เนื้อความทั้งหมดจากแฟ้ม `p4chap2.xml` จะถูกนำไปแทนที่ entity ChapTwo นั้น

นอกจากการกำหนด entity เพื่อใช้ในตัวเอกสาร XML แล้ว ยังสามารถกำหนด entity สำหรับใช้ในส่วน DTD ได้ด้วย แต่การกำหนด entity เพื่อใช้ใน

แฟ้มข้อมูล DTD จะมีข้อแตกต่างเพิ่มเติม คือ ต้องเติม % หลังคำว่า entity และในเวลาที่ใช้ entity ประเภทนี้ให้ใช้เครื่องหมาย % แทน & ดังตัวอย่างนี้

```
<!ENTITY % TEI.prose 'INCLUDE'>
<!ENTITY % TEI.extensions.dtd SYSTEM 'mystuff.dtd'>
```

ในตัวอย่างนี้ เมื่อมีการอ้างถึง entity TEI.prose ภายในส่วนของ DTD entity นี้จะถูกแทนด้วยคำว่า INCLUDE หรือเมื่อมีการอ้าง entity TEI.extensions.dtd ภายในส่วนของ DTD entity นี้ก็จะถูกแทนที่ด้วยเนื้อความจากแฟ้มข้อมูล mystuff.dtd entity ประเภทหลังนี้เรียกว่า parameter entity เมื่ออ้างถึงต้องใช้เครื่องหมาย % นำหน้าแทนที่จะใช้ & แต่ยังปิดท้ายด้วยเครื่องหมาย ; เหมือนเดิม

5. NOTATION ใช้เพื่อกำหนดว่าจะอนุญาติให้มีข้อมูลที่ไม่ใช่ XML อยู่ภายในเอกสารนั้นได้หรือไม่ ตัวอย่างเช่น

```
<!ENTITY fig1 SYSTEM "figure1.png" NDATA png>
<!NOTATION png PUBLIC
    '-//TEI//NOTATION IETF RFC2083 Portable Network
Graphics//EN'>
```

ใช้เพื่อกำหนดว่าจะยอมให้มีการใช้รูปแบบ (notation) ของรูปภาพที่เป็น png ปรากฏอยู่ในเอกสารได้ด้วยโดยอ้างด้วย entity &fig1;

4.1.4.2 ตัวเชื่อม (Connectors)

Connectors คือ ตัวเชื่อมที่ใช้บอกความสัมพันธ์ของการเกิดร่วมกันของหน่วยต่างๆ ตัวเชื่อมที่กำหนดให้ใช้ได้ มีสามตัวดังนี้

```
,       SEQ     all must occur in the order specified
&       AND     all must occur in any order
|       OR      one and only one must occur
```

กล่าวคือ เครื่องหมาย , ใช้บอกลำดับของหน่วยว่ามีอะไรบ้างและระบุว่าอะไรปรากฏก่อน อะไรปรากฏหลัง ส่วน & มีความหมายเหมือนกับ , เพียงแต่ไม่ได้ระบุลำดับการปรากฏที่ตายตัว หน่วยใดจะปรากฏก่อนหรือหลังก็ได้ ส่วน |

ใช้กำหนดให้เลือกเพียงหน่วยใดหน่วยหนึ่งเท่านั้น นอกจากนี้ ยังมีเครื่องหมายที่ใช้บอกความถี่ของการปรากฏ (occurrence indicator) อีกสามตัว ได้แก่

```
+       PLUS    required and repeatable (1 or more times)
?       OPT     optional (0 or 1 time)
*       REP     optional and repeatable (0 or more times)
```

กล่าวคือ เครื่องหมาย + ที่ตามหลังหน่วยใดใช้เพื่อแสดงว่าหน่วยนั้นเป็นหน่วยบังคับและอาจปรากฏมากกว่าหนึ่งครั้งก็ได้ ส่วน ? บอกว่าหน่วยนั้นไม่บังคับ จะปรากฏหนึ่งครั้งหรือจะไม่ปรากฏก็ได้ ส่วน * นั้นคล้ายกับ ? คือไม่บังคับ แต่ต่างกันตรงที่อาจปรากฏได้มากกว่าหนึ่งครั้งได้

4.1.4.3 การกำหนดชนิดเนื้อหา (Content Types)

คือการกำหนดถึงเนื้อหาของแต่ละหน่วยว่าเป็นข้อมูลชนิดไหน ดังตัวอย่างต่อไปนี้

- PCDATA (parsed character data) บอกว่าให้ดูเฉพาะตัวอักษรปกติที่รู้จัก ส่วนเครื่องหมายเฉพาะ เช่น < > จะเป็นส่วนหนึ่งของแท็กจะไม่ถือว่าเป็นส่วนหนึ่งของข้อมูล เป็นต้น

- CDATA (character data) บอกว่าข้อมูลเป็นตัวอักษรปกติ เครื่องหมายเฉพาะจำพวก < > หรือ & ก็จะถือว่าเป็นส่วนหนึ่งของข้อมูลด้วย

- RCDATA (replaceable character data) จะเหมือนกับ CDATA ยกเว้นในส่วนที่ขึ้นต้นด้วย & จะมองว่าเป็นการบอกถึง entity เช่น <a &subk;>&sub+; จะหมายถึง $<a_k>_+$

- EMPTYบอกว่าหน่วยนั้นไม่มีเนื้อความหรือข้อมูล

4.2 การกำหนดส่วนเฉพาะ (Marked Section)

เป็นวิธีการกำกับ (mark) บางส่วนของเอกสารเพื่อวัตถุประสงค์เฉพาะ ซึ่งมีใช้ในสองกรณี กรณีแรกเป็นการกำหนดส่วนเฉพาะเพื่อให้ไม่มองข้อมูลเป็นแบบ XML ให้มองเป็นสายอักขระตามที่เห็น เช่น

```
<p>The <gi>term</gi> element may be used to mark any
  technical term:
  <eg><![CDATA[
  This <term>recursion</term> is giving me a headache.
  ]]></eg></p>
```

เป็นการบอกว่าแท็ก <term> ที่ปรากฏไม่ใช่การกำกับของ XML แต่เป็น CDATA กรณีที่สองเป็นการกำกับส่วนเฉพาะที่แปรไปตามเงื่อนไข (conditional marked section) ว่าจะต้องการให้ข้อความนั้นปรากฏหรือไม่ปรากฏ เช่น ในกรณีข้อความทางกฎหมายซึ่งอาจแตกต่างกันไปตามแต่ละท้องถิ่น สามารถกำหนด marked section ให้ include หรือ ignore ข้อความบางตอนเพื่อให้เหมาะสมกับกฎหมายในท้องถิ่นนั้น เช่น ข้อความว่า “Liability is limited to $50,000.” ไม่ควรปรากฏหากจัดพิมพ์ในรัฐ Delaware แต่ควรจะปรากฏหากจัดพิมพ์ในรัฐ Maryland ตัวอย่างเช่น หากภายในเอกสารมีข้อความ

```
In such cases, the bank will reimburse the customer for
all losses.  <![ %Maryland; [ Liability is limited to
$50,000. ]]>
```

ถ้า entity Maryland ถูกกำหนดใน DTD ว่าเป็น INCLUDE ดังนี้

```
<!ENTITY % Maryland 'INCLUDE'>
```

ข้อความ Liability is limited to $50,000. ก็จะปรากฏออกมาด้วย แต่ถ้า entity Maryland ถูกกำหนดเป็น IGNORE ข้อความนี้ก็จะไม่ปรากฏ

4.3 ตัวอย่างของแฟ้มข้อมูลที่กำกับด้วย SGML และ XML

แฟ้มข้อมูลที่กำกับตามมาตรฐานของ SGML จะประกอบด้วยส่วนต้นของเอกสาร (document prolog) และส่วนของตัวเอกสาร (document instance) ส่วนต้นของเอกสารจะประกอบด้วย SGML Declaration และส่วน DTD แต่เนื่องจากส่วน SGML Declaration เป็นการกำหนดข้อมูลพื้นฐานของ SGML ที่ใช้ ผู้ใช้ทั่วไปจึงไม่จำเป็นต้องระบุส่วนนี้ ผู้ใช้ทั่วไปจะระบุเฉพาะส่วนของ DTD ซึ่งจะถูกนำมาใช้เพื่อตรวจสอบเอกสารนั้น ๆ ว่ามีโครงสร้างตรงตามที่กำหนดไว้ใน DTD หรือไม่ (valid)

ลักษณะของแฟ้มข้อมูลที่กำกับด้วย SGML จึงขึ้นต้นด้วย <!DOCTYPE ซึ่งจะกำหนด DTD ที่ใช้ ส่วนของ DTD นี้มักจะอ้างอิงแฟ้มข้อมูล DTD ที่ตั้งขึ้นเป็นมาตรฐานสำหรับเอกสารประเภทต่าง ๆ สำหรับข้อกำหนดอื่น ๆ ที่ใช้เฉพาะในเอกสารนั้น จะระบุเพิ่มเติมไว้ภายในเอกสารนั้น เช่น ในรูป 10.4 tei2.dtd เป็นแฟ้มข้อมูลที่กำหนด DTD ทั่วไปสำหรับเอกสารประเภทนี้ทั้งหมด ส่วนของ !ENTITY และ !ELEMENT ในบรรทัดถัดมาเป็นการกำหนดใช้เฉพาะสำหรับเอกสารชิ้นนี้เท่านั้น

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P3//DTD Main Document
Type//EN" "tei2.dtd" [
  <!ENTITY tla "Three Letter Acronym">
  <!ELEMENT myTag (#PCDATA)    >
  <!-- any other special-purpose declarations or
       re-declarations go  here -->
  ]>
<TEI.2>
<!-- This is an instance of a modified TEI.2 type
document, which may contain <myTag>my special
tags</myTag> and references to my usual entities such as
&tla;. -->
</TEI.2>                                    (TEI P4: Chapter 2)
```

รูป 10.4 ตัวอย่างแฟ้มข้อมูลที่กำกับตามมาตรฐาน XML

ถัดจากส่วนของ DTD ก็จะเป็นส่วนของ document instance ซึ่งเป็นส่วนของข้อมูลเอกสารที่กำกับด้วยแท็กต่างๆ ตามลักษณะที่กำหนดไว้ใน DTD ส่วนของ document instance นี้อาจจะพิมพ์ต่อจากส่วนของ DTD ไว้ในแฟ้มข้อมูลนั้นเลยก็ได้ ดังแสดงในรูปที่ 10.4 แต่โดยปกติแล้ว ส่วนของ document instance นี้มักจะมีขนาดใหญ่ จึงมักแยกออกมาเป็นแฟ้มข้อมูลต่างหาก ดังในตัวอย่างรูปที่ 10.5 ซึ่ง document instance ที่อยู่ในแฟ้มข้อมูลเดียวกับ DTD มีเฉพาะส่วนที่เป็น header ที่บอกข้อมูลที่มาของเอกสาร แต่ในส่วนเอกสารทั้งหมดซึ่งเป็นข้อความในหนังสือเรื่อง The Last of Mohicans ถูกแยกไว้ในอีกแฟ้มข้อมูลตามที่ระบุไว้ใน entity ที่ชื่อ file1 (ดูที่บรรทัดก่อนสุดท้ายและบรรทัดที่ 4 ตามลำดับ) ซึ่งมีชื่อว่า mohicans.1976

```
<!DOCTYPE ota SYSTEM "ota.dtd" [
<!ENTITY % OTAents system "unixlat0.dtd">
%OTAents
<!ENTITY file1 SYSTEM "mohicans.1976">
]>
<ota n="1976" crdate="1993-08-10" update="1993-11-01">
<header>
<fileDesc>
<titlStmt>
<title>Cooper's "The Last of the Mohicans": electronic
edition</title>
<edStmt>
      Public Domain TEI edition prepared at the Oxford
Text Archive
<extent>
      Filesize uncompressed: 893 Kbytes.
<pubStmt>
<resp>
<role>Distributors</role>
<name>Oxford Text Archive,
     Oxford University Computing Services,
     13 Banbury Road, Oxford OX2 6NN;
     archive@ox.ac.uk
</name>
</resp>
<idno>1976</idno>
```

```
<avail><p>Freely available for non-commercial
use provided that this header is included in its
entirety with any copy distributed
</p>
</avail>
<date>10 Aug 1993</date>
</pubStmt>
<!-- seriesstmt -->
<noteStmt>
<note>This is a prototype header</note>
</noteStmt>
<srcDesc>
<p>First edition published in 1826. Keyed into
electronic form
by Joanne Hindman, OED North American Reading Program.
</p>
</srcDesc></fileDesc>
<encDesc>
<editDecl>
<p>
</p>
<refsDecl>
<p>Chapters (div) bear n attributes in the form C1.
</p>
</refsDecl>
</encDesc>
<revDesc>
<change><date>Aug 93</date>
<resp><role>Check
text</role><name>JAT</name></resp></change>
</revDesc>
</header>
&file1;
</ota>
```

รูป 10.5 แฟ้มข้อมูลหนังสือเรื่อง The Last of Mohicans ซึ่งกำกับตาม SGML (จาก Oxford Text Arichive)

ส่วนแฟ้มข้อมูล XML นั้นจะมีลักษณะเดียวกันกับแฟ้มข้อมูล SGML ที่กล่าวมาคือมีส่วนเริ่มต้น (prolog) ที่ประกอบด้วย XML declaration และส่วน DTD จากนั้นจึงเป็นส่วนเนื้อหาหรือ document instance ส่วน XML

declaration มีเพียงหนึ่งบรรทัดซึ่งเรียกว่าเป็น processing instruction แล้วตามด้วย document type definition c]t document instance ดังตัวอย่างข้างล่าง

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE myDoc SYSTEM "myDoc.dtd" [
  <!ENTITY tla "three letter acronym">]>
<myDoc>
  <p>This is an instance of a "myDoc" document.</p>
  <p>It is surprisingly free of &tla;s.</p>
</myDoc>
```

รูป 10.6 แฟ้มข้อมูลกำกับแบบ XML (TEI P4 Chapter 2)

5. แท็กพื้นฐานที่กำหนดโดย TEI

เนื่องจาก TEI เป็นผู้ที่กำหนดมาตรฐานของรายการแท็ก และ DTD สำหรับใช้ในการกำกับข้อมูลเอกสารอิเล็กทรอนิกส์แบบต่าง ๆ ผู้ที่จะสร้างคลังข้อมูลภาษาโดยที่มีการกำกับข้อมูลด้วย จึงจำเป็นต้องศึกษาหนังสือคู่มือแนวทางการกำกับข้อมูลที่ออกโดย TEI นี้ แต่เนื่องจากคู่มือดังกล่าวมีจำนวนทั้งสิ้น 39 บทมีความยาวมากกว่า 2,000 หน้า จึงเป็นการยากสำหรับผู้เริ่มต้นศึกษา จึงมีการจัดทำหนังสือคู่มือฉบับย่อหรือ TEI Lite สำหรับผู้เริ่มต้นศึกษา โดยรวบรวมรายการแท็กที่ผู้ใช้ทุกคนควรจะรู้ไว้เป็นกลุ่มของแท็กเริ่มต้น (starter set) ในการทำความเข้าใจกับข้อกำหนดของ TEI คู่มือนี้คือ TEI Lite: An Introduction to Text Encoding for Interchange หรือมีชื่อย่อว่า TEI U5 (http://www.tei-c.org/Guidelines/Customization/Lite/) ตัวอย่างโครงการที่มีการกำกับข้อมูลโดยใช้ TEI Lite ได้แก่ Oxford Text Archive, โครงการ Electronic Text Centers ที่จัดทำที่ University of Virginia และที่จัดทำที่ University of Michigan

5.1 โครงสร้างของเอกสารตามข้อกำหนด TEI

เอกสารที่ได้มาตรฐานของ TEI จะต้องประกอบด้วย สองส่วน คือมีส่วนที่เป็นส่วนหัว (header) กับ ส่วนที่เป็นตัวบท (text) ส่วน header ให้ข้อมูลเกี่ยวกับที่มาและการจัดสร้างเอกสารนั้น ส่วนตัวบทเป็นเนื้อความของเอกสาร

โครงสร้างของเอกสารเดี่ยว เมื่อถูกกำกับข้อมูล (encode) ตามที่ TEI กำหนดแล้วจะเป็นดัง

```
<TEI.2>
    <teiHeader> [ TEI Header information ]  </teiHeader>
    <text>
        <front> [ front matter ... ]   </front>
        <body>  [ body of text ... ]   </body>
        <back>  [ back matter ...  ]   </back>
    </text>
</TEI.2>
```

รูป 10.7 โครงสร้างเอกสารเดี่ยวตามคำแนะนำของ TEI

ส่วนข้อมูลที่ประกอบด้วยตัวบท (text) หลายๆบท สามารถที่จะกำหนดส่วนที่เป็นส่วนหน้า (front) และส่วนหลัง (back) แยกเฉพาะของแต่ละตัวบทได้ และก็สามารถกำหนดส่วนหน้าและหลังรวมของตัวบททั้งหมดได้ โครงสร้างของเอกสารที่ TEI กำหนดเอาไว้เป็นดังนี้

```
<TEI.2>
  <teiHeader> [ header information for the composite ]
</teiHeader>
  <text>
   <front> [ front matter for the composite ]</front>
    <group>
     <text>
      <front> [ front matter of first text ] </front>
      <body>  [ body of first text  ] </body>
      <back>  [ back matter of first text ]</back>
     </text>
     <text>
      <front> [ front matter of second text]  </front>
      <body>  [ body of second text  ]          </body>
```

```
      <back>  [ back matter of second text ]    </back>
     </text>
       [ more texts or groups of texts here ]
     </group>
   <back>  [ back matter for the composite  ]   </back>
  </text>
</TEI.2>
```

รูป 10.8 โครงสร้างเอกสารกรณีที่มีเอกสารหลาย ๆ เล่มในชุด

เมื่อนำเอกสารตามแบบ TEI หลายๆ เอกสารมารวมกันเป็นคลังข้อมูลภาษา แต่ละเอกสารจะมีส่วนหัว (header) ของตัวเอง และทั้งหมดจะอยู่ภายใต้แท็กที่กำหนดขอบเขตคลังข้อมูลภาษา เช่น "teiCorpus" ดังแสดงในรูปที่ 10.9 ซึ่งโครงสร้างนี้แสดงภาพใหญ่ของคลังข้อมูลภาษาว่าประกอบด้วยรายการเอกสารอะไรบ้าง

```
<teiCorpus>
  <teiHeader>[header information for the corpus]
  </teiHeader>
  <TEI.2>
     <teiHeader>[header information for first text]
     </teiHeader>
     <text> [first text in corpus]   </text>
  </TEI.2>
  <TEI.2>
    <teiHeader>[header information for second text]
    </teiHeader>
    <text>   [second text in corpus]   </text>
  </TEI.2>
</teiCorpus>
```

รูป 10.9 โครงสร้างของคลังข้อมูลภาษาซึ่งรวมเอกสารหลาย ๆ เอกสาร

5.2 การกำกับข้อมูลส่วนเนื้อความ

TEI ได้กำหนดแท็กพื้นฐานต่าง ๆ ที่ใช้ในการกำกับข้อมูลเอกสาร ตัวอย่างต่อไปนี้ แสดงถึงแท็กพื้นฐานต่าง ๆ

- <front> ใช้กำกับข้อมูลส่วนหน้าของเอกสารซึ่งปรากฏก่อนส่วนเนื้อความ เช่น หัวเรื่อง (headers) หน้าปก (title page) หน้าบทนำ (preface) หน้าคำอุทิศ (dedication) เป็นต้น
- <group> ใช้กำกับกลุ่มข้อมูลเอกสารว่าเป็นเอกสารเดี่ยวหรือเป็นเอกสารกลุ่ม
- <body> ใช้บอกส่วนเนื้อความทั้งหมดของเอกสารแต่ละชิ้น โดยไม่รวมส่วนที่เป็นส่วนหน้าและส่วนท้าย
- <back> ใช้กำกับข้อมูลส่วนท้ายของเอกสาร เช่น บรรณานุกรม

ในส่วนตัวบทของเอกสารอาจประกอบด้วยข้อความหลายๆ ย่อหน้า (paragraph) ซึ่งหลายๆ ย่อหน้าอาจรวมเป็นบท (chapter) เป็นตอน (sections) เป็นต้น ซึ่งแต่ละย่อหน้าจะกำกับด้วยแท็ก <p> ส่วนการกำกับบทหรือตอนนั้นจะใช้แท็ก หรือ <div1> สำหรับแต่ละส่วนได้ ดังนี้

- <p> ใช้กำกับข้อความแต่ละย่อหน้า
- ภายในจะประกอบด้วยส่วนย่อย (subdivision) ของส่วนหน้า ส่วนเนื้อความ หรือส่วนท้ายของตัวบท
- <div1> เป็นส่วนที่ย่อยลงมาจาก ถ้าต้องการแยกย่อยลงไปอีกก็จะมี <div2> <div3> ลงไปเรื่อยๆ จนถึง <div7> หากจำเป็นต้องใช้มากกว่า <div7> ก็จะต้องแก้ไขข้อกำหนดใน DTD หรืออาจใช้ เฉยๆ เพราะสามารถกำหนดให้มี ย่อยภายในได้

แท็กทั้งสามนี้มีคุณสมบัติ (attributes) เหมือนกัน ได้แก่

- type ใช้บอกหมวดหมู่ของเนื้อหาส่วนนั้นว่าเป็นอะไร เช่น "Book", "Chapter", "Poem", เป็นต้น หรือเป็น "Group" สำหรับกลุ่มของบทกวีนิพนธ์ หรือเป็น "Sonnet", "Speech", และ "Song" ซึ่งค่าที่กำหนดนี้จะถือเป็นการบอกลักษณะของข้อมูลส่วนนั้นไปจนกว่าจะมีการกำหนดค่า type ใหม่สำหรับ ต่อไป

- id ใช้บอกชื่อของเนื้อหาแต่ละส่วน ซึ่งต้องเป็นชื่อที่ไม่ซ้ำกัน ชื่อ id นี้สามารถอ้างอิงในการทำ cross references ได้
- n ใช้บอกชื่อย่อของแต่ละส่วน โดยมากเป็นไปตามที่ปรากฏอยู่ในเอกสารนั้นๆ แล้ว เช่น “1.1”, “1.2” เป็นต้น

ตัวอย่างข้างล่างเป็นการใช้แท็กและคุณลักษณ์ดังกล่าว

```
<div1 id=WN1 n='I' type='book'>
  <div2 id=WN101 n='I.1' type='chapter'> ... </div2>
  <div2 id=WN102 n='I.2' type='chapter'> ... </div2>
  ...
  <div2 id=WN110 n='I.10' type='chapter'>
     <div3 id=WN1101 n='I.10.1' type=part> ... </div3>
     <div3 id=WN1102 n='I.10.2' type=part> ... </div3>
  </div2>
  ...
</div1>
<div1 id=WN2 n='II' type='book'>
   ....
</div1>
...
```

รูป 10.10 ตัวอย่างการใช้แท็กและกำหนดคุณสมบัติภายในแท็ก div

สำหรับรายละเอียดของแท็กในส่วนอื่นๆ เช่น แท็กย่อยที่ใช้กำกับข้อมูลส่วนเนื้อหา แท็กที่ใช้ในการกำกับข้อมูลแบบต่างๆ เช่น ตาราง รูปภาพ วันที่ บรรณานุกรม ฯลฯ ผู้สนใจสามารถศึกษาได้จากคู่มือ TEI Lite: An Introduction to Text Encoding for Interchange ซึ่งจะให้ข้อมูลของแท็กพื้นฐานที่ผู้จัดทำคลังข้อมูลภาษาหรือผู้ใช้ควรจะรู้

5.3 การกำกับข้อมูลในคลังข้อมูลภาษา

นอกจากคู่มือ TEI Lite ซึ่งให้ข้อมูลพื้นฐานสำหรับการกำกับข้อมูลเอกสารทั่วๆไปแล้ว สำหรับผู้ที่จะสร้างคลังข้อมูลภาษาควรจะได้ศึกษารายงานของ TEI บทที่ 23 ด้วยเพราะเป็นบทที่กล่าวถึงการกำกับข้อมูลสำหรับคลังข้อมูล

ภาษาโดยเฉพาะ ดังนั้นผู้ที่จะกำกับคลังข้อมูลภาษาจึงควรจะศึกษาคู่มือตามคำแนะนำของ TEI ในส่วนที่เกี่ยวข้อง แล้วจึงพัฒนาภาษาสำหรับกำกับข้อมูล (markup language) ที่จะใช้ให้สอดคล้องกับ SGML และมีรายการแท็กที่มีคุณสมบัติตามที่ TEI กำหนดไว้

ภาษาที่ใช้กำกับข้อมูลในคลังข้อมูล BNC เป็นตัวอย่างหนึ่งของการใช้ SGML ในการกำกับคลังข้อมูลภาษา BNC กำหนดให้ CDIF (Corpus Document Interchange Format) เป็นภาษาสำหรับกำกับข้อมูล CDIF เป็นการนำ SGML มาประยุกต์ใช้ โดยให้สอดคล้องกับคำแนะนำของ TEI ที่กำหนดไว้ในเรื่องของการกำกับข้อมูลทางภาษาศาสตร์ ซึ่งต่อมา BNC ได้ปรับข้อมูลทั้งหมดจากมาตรฐาน SGML มาเป็น XML แล้วและจำหน่ายในชื่อของ BNC XML Edition

CES เป็นอีกตัวอย่างหนึ่งของการนำ SGML มาประยุกต์ใช้ CES (Corpus Encoding Standard) เป็นความร่วมมือระหว่าง MULTEXT (Multilingual Text Tools and Corpora), EAGLES (Expert Advisory Group on Language Engineering Standards) และ Vassar/CNRS (supported by the U.S. National Science Foundation) โดยมีจุดมุ่งหมายที่จะกำหนดมาตรฐานสำหรับการกำกับข้อมูล (encode) ทางด้านวิศวกรรมภาษา (language engineering) CES จะกำหนดมาตรฐานสำหรับกำกับข้อมูลทางภาษาศาสตร์ไว้โดยละเอียด ผู้ที่จะสร้างคลังข้อมูลภาษาเพื่องานทางภาษาศาสตร์ จึงควรศึกษารายงานของ CES โดยละเอียด ตัวอย่างของเอกสารคู่มือที่จัดทำโดย EAGLES ที่เผยแพร่ทางอินเทอร์เน็ต ได้แก่

- Preliminary recommendations on corpus typology
- Preliminary Recommendations on Text Typology

- Recommendations on corpus encoding
- Recommendations for the Morphosyntactic Annotation of Corpora
- Preliminary Recommendations for the Syntactic Annotation of Corpora (ในรายงานมีสรุปภาพรวมของการกำกับข้อมูลทางวากยสัมพันธ์ (syntactic annotation) 6 แบบที่ใช้กันอยู่ คือ SUSANNE, ENGCG, TOSCA, UPenn, Lancaster/IBM, และ IBM Paris Treebank)
- Preliminary Recommendations on subcatgorization
- Preliminary recommendations on spoken text corpora
- Preliminary recommendations on Lexical Semantic Encoding

(ดู http://www.ilc.cnr.it/EAGLES96/home.html)

โครงการ CES ในปัจจุบันก็เปลี่ยนไปใช้มาตรฐาน XML โดยเปลี่ยนชื่อโครงการเป็น XCES และกำลังอยู่ระหว่างการปรับเอกสารต่างๆ ที่ทำไว้ให้สอดคล้องกับคู่มือ TEI P5 ผู้สนใจสามารถตรวจสอบข้อมูลใหม่ๆ ได้ที่ http://www.xces.org/

นอกจาก CDIF และ CES แล้ว SGML/XML ยังได้รับการยอมรับและใช้ในการออกแบบภาษาสำหรับกำกับข้อมูล (markup language) ในโครงการต่างๆมากมาย ซึ่งมีทั้งที่ใช้รายการแท็กพื้นฐานตามที่ TEI แนะนำและที่เพิ่มเติมแท็กอื่นๆ ตามความเหมาะสม เช่น Oxford Text Archive (OTA), University of Virginia Electronic Text Center, The Electronic Archive of Early American Fiction (UVA), University of Michigan Humanities Text Initiative (HTI), HTI American Verse Project, Model Editions Partnership: Historical Editions in the Digital Age, American Memory Project, Library of Congress, Brown University Scholarly Technology Group (STG), Midrash Pirqe Rabbi

Eliezer Project (Hebrew Union College), University of Cincinnati College of Law, Center for Electronic Text in the Law, Linguistic Data Consortium (LDC), IATH - Piers Plowman Archive, IATH - Rossetti Archive, Japanese Text Initiative (University of Virginia and the University of Pittsburgh), Center for Electronic Texts in the Humanities (CETH), The Canterbury Tales Project, University of Pittsburgh Electronic Text Project, Georgetown University: Labyrinth Medieval Studies and Peirce Projects, European Corpus Initiative (ECI), Language Technology Group (LTG), Human Communication Research Centre (HCRC), University of Edinburgh, Lingua Parallel Concordancing Project, University of Waterloo Centre for the New OED and Text Research เป็นต้น

(ดู http://xml.coverpages.org//sgml.html)

5.4 การกำกับคลังข้อมูลภาษาตามแนวทางของ TEI

รายงานของ TEI บทที่ 23 เรื่อง Language Corpora เป็นบทที่แนะนำแนวทางการกำกับคลังข้อมูลภาษา ซึ่งนอกจากรายการแท็กที่กล่าวไว้ในบทนี้ ผู้อ่านจำเป็นต้องอ่านรายงานบางบทก่อนหน้านี้ที่กำหนดรายการแท็กพื้นฐานที่ใช้ร่วมกัน เช่น ในบทที่ 6 เรื่อง Elements Available in All TEI documents บทที่ 7 เรื่อง Default Text Structure, เป็นต้น รวมทั้งบทที่ 15 เรื่อง Simple Analytic Mechanisms ซึ่งมีการกล่าวถึงการกำกับข้อมูลทางภาษาศาสตร์

5.4.1 การกำกับข้อมูลบริบท

รายการแท็กในบทที่ 23 นี้ให้ข้อมูลต่างๆ เกี่ยวกับบริบท (contextual information) เช่น อายุ เพศ ถิ่นกำเนิดของผู้พูดในข้อมูล สถานภาพทางสังคม

เศรษฐกิจ ข้อมูลบรรณานุกรม หัวเรื่อง ทำเนียบภาษา (register) เป็นต้น โครงสร้างหลักประกอบด้วยสามส่วน คือ <textDesc> <particDesc> และ <settingDesc>

1. <textDesc> เป็นแท็กที่ใช้กำกับข้อมูลเกี่ยวกับสถานการณ์ที่การใช้ตัวบท (text) นั้นเกิดขึ้น ภายในประกอบด้วยแท็กย่อยต่างๆ เช่น

- <channel> : ใช้กำกับข้อมูลสื่อหรือช่องทางที่ตัวบทนั้นถูกนำเสนอออกมา เช่น ในตัวบทที่เป็นภาษาเขียน channel อาจเป็นสิ่งพิมพ์ ต้นฉบับที่เขียนเป็นลายมือ (manuscript) หรือ อีเมล์ เป็นต้น ในตัวบทที่เป็นภาษาพูด อาจเป็นวิทยุโทรศัพท์ หรือแบบเผชิญหน้า (face-to-face) คุณสมบัติที่สามารถกำหนดในแท็กนี้ เช่น

  mode : ใช้บอกโหมดของการใช้ภาษาผ่านสื่อว่าเป็นภาษาพูดหรือภาษาเขียน ค่าที่สามารถกำหนดใช้ได้ มีดังนี้

  - s : เป็นภาษาพูด
  - w : เป็นภาษาเขียน
  - sw : เป็นภาษาพูดสำหรับเขียน (เช่น การบอกจด dictation)
  - ws : เป็นภาษาเขียนขึ้นสำหรับพูด (เช่น บทสนทนา)
  - m : โหมดผสม
  - x : ไม่ทราบโหมด

- <constitution> : ใช้อธิบายองค์ประกอบภายในตัวบทว่ามีลักษณะอย่างไร สมบูรณ์ หรือมีเพียงบางส่วน คุณสมบัติที่กำหนดได้มีดังนี้

  type : ใช้กำหนดว่าตัวบทนั้นถูกประกอบขึ้นมาอย่างไร ค่าที่กำหนดใช้ได้มี

  - single : เป็นตัวบทเดี่ยวที่สมบูรณ์

- composite : เป็นตัวบทที่ประกอบด้วยตัวบทย่อยหลายส่วน แต่ละส่วนมีครบสมบูรณ์
- frags : เป็นตัวบทที่ประกอบด้วยตัวบทย่อยหลายส่วน แต่ละส่วนไม่จำเป็นต้องมีครบสมบูรณ์
- unknown : ไม่ทราบลักษณะองค์ประกอบ

- <derivation> : ใช้อธิบายความเกี่ยวข้องกับตัวบทอื่น คุณสมบัติที่กำหนดได้มีดังนี้

type : ใช้กำหนดประเภทที่มาของตัวบทนั้น ตัวอย่างค่าที่ใช้ได้มี

- original : เป็นตัวบทต้นฉบับ
- revision : เป็นตัวบทที่แปลงแก้ไขมาจากตัวบทอื่น
- translation : เป็นตัวบทแปล
- abridgment : เป็นตัวบทที่เป็นฉบับตัดตอนหรือฉบับย่อของตัวบทอื่น
- plagiarism : เป็นตัวบทที่แอบคัดลอกจากตัวบทอื่น (plagiarized)
- traditional : ไม่มีที่มาชัดเจน แต่เป็นตัวบทที่ได้เค้ามาจากตัวบทโบราณ

- <domain> : ใช้อธิบายบริบททางสังคมที่มีในการใช้ข้อความนั้น เช่น ใช้ส่วนตัว ใช้ในที่สาธารณะ ใช้ในการศึกษา ศาสนา เป็นต้น คุณสมบัติที่กำหนดได้มีดังนี้

type : ใช้กำหนดขอบเขตของการใช้ ค่าที่สามารถกำหนดได้ เช่น

- art : เป็นเรื่องทางศิลปะและบันเทิง
- domestic : เป็นเรื่องในครอบครัวหรือส่วนตัว
- religious : เป็นเรื่องทางศาสนาหรือพิธีกรรม
- business : เป็นเรื่องทางธุรกิจหรือสถานที่ทำงาน

- education : เป็นเรื่องการศึกษา
- govt : เป็นเรื่องทางรัฐบาลหรือกฎหมาย
- public : เป็นเรื่องของสาธารณะ

- <factuality> : ใช้อธิบายลักษณะของตัวบทว่าเป็นเรื่องแต่งหรือไม่ คุณสมบัติที่กำหนดได้มี

type : ใช้บอกประเภทของตัวบทนั้น ค่าที่กำหนดได้มีดังนี้

- fiction : ทั้งหมดถือเป็นเรื่องแต่ง
- fact : ทั้งหมดถือเป็นเรื่องจริงหรือข้อเท็จจริง
- mixed : มีลักษณะผสมทั้งเรื่องแต่งและเรื่องจริง
- inapplicable : ไม่สามารถแยกตามเกณฑ์นี้ได้

- <interaction> : ใช้อธิบายลักษณะการปฏิสัมพันธ์ระหว่างผู้สร้างตัวบทและผู้รับสาร คุณสมบัติที่กำหนดได้มี

type : ใช้ระบุว่ามีปฏิสัมพันธ์ระหว่างผู้ร่วมในเหตุการณ์หรือไม่ ค่าที่กำหนดใช้ได้มี

- none : ไม่มีปฏิสัมพันธ์ใดใด เช่น เป็น บทพูดคนเดียว
- partial : มีปฏิสัมพันธ์บางส่วน เช่น เป็นบทพูดของคนเดียว แต่อาจมีการตอบสนองจากผู้ฟังเป็นครั้งคราว
- complete : มีปฏิสัมพันธ์เต็มที่ เช่น เป็นการสนทนาแบบเผชิญหน้า
- inapplicable : ไม่สามารถที่จะระบุลักษณะแบบนี้ได้หรือไม่เหมาะสมที่จะระบุลักษณะแบบนี้

active : ใช้ระบุจำนวนผู้มีส่วนร่วมในในการสร้างตัวบทนั้น ค่าที่กำหนดได้มีดังนี้

- singular : มีคนเดียว

- plural : มีหลายคน

- corporate : เป็นในนามของบริษัท องค์กร

- unknown : ไม่ทราบจำนวน

passive : ใช้ระบุจำนวนผู้ที่เป็นผู้รับสารของตัวบทนั้น ค่าที่กำหนดได้มีดังนี้

- self : เป็นการสื่อกับตัวเอง

- single : เป็นตัวบทที่สื่อกับบุคคลอื่นคนเดียว เช่น จดหมาย

- many : เป็นตัวบทที่สื่อกับบุคคลอื่นมากกว่าหนึ่งคนและทราบจำนวนชัดเจน เช่น การสนทนากลุ่ม

- group : เป็นตัวบทที่สื่อกับบุคคลอื่นมากกว่าหนึ่งคนแต่ไม่ทราบจำนวนชัดเจน เช่น การบรรยายในชั้นเรียน

- world : เป็นตัวบทที่สื่อกับคนจำนวนมาก ไม่ระบุชัดเจน เช่น หนังสือ

- <preparedness> : ใช้อธิบายลักษณะว่าตัวบทนั้นมีการเตรียมการหรือไม่ คุณสมบัติที่กำหนดได้มี

type : ใช้ระบุลักษณะของการเตรียมการ ตัวอย่างค่าที่เป็นไปได้ เช่น

- none : ไม่มีการเตรียมการ

- scripted : มีบทล่วงหน้า

- formulaic : เป็นไปตามรูปแบบที่รู้กัน

- revised : มีการตรวจแก้ไขก่อนนำเสนอ

- <purpose> : ใช้บอกถึงจุดมุ่งหมายของการใช้ตัวบทนั้นในการสื่อสาร คุณสมบัติที่กำหนดได้มีดังนี้

type : ใช้ระบุวัตถุประสงค์ ค่าที่กำหนดได้ คือ

- persuade : เพื่อโฆษณาชวนเชื่อ

- express : เพื่อบรรยายตัวเองหรือสารภาพ

- inform : เพื่อให้ข้อมูลข่าวสารหรือความรู้

- entertain : เพื่อความบันเทิง

degree : ใช้ระบุระดับของวัตถุประสงค์ ค่าที่ระบุได้คือ

- high : มีระดับสูง

- medium : มีในระดับปานกลาง

- low : มีในระดับต่ำ

- unknown : ไม่ทราบ

ตัวอย่างที่ปรากฏข้างล่างเป็นตัวอย่างของ <textDesc> ที่กำกับข้อมูลที่เป็นการสนทนาแบบไม่เป็นทางการในบ้าน และ <textDesc> ที่กำกับข้อมูลที่เป็นนวนิยาย ตามลำดับ

```
<textDesc id=t1 n='Informal domestic conversation'>
   <channel mode=s>informal face-to-face
conversation</channel>
   <constitution type=single>each text represents a
continuously recorded interaction among the specified
participants  </constitution>
  <derivation type=original>
  <domain type=domestic>plans for coming week, local
affairs</domain>
  <factuality type=mixed>mostly factual, some
jokes</factuality>
  <interaction type=complete active=plural passive=many>
       <preparedness type='spontaneous'>
       <purpose type=entertain degree=high>
       <purpose type=inform degree=medium>
</textDesc>
```

รูป 10.11 ตัวอย่าง <textDesc> ของการสนทนาในบ้าน

```
<textDesc n='novel'>
          <channel mode=w>print; part issues</channel>
          <constitution type=single>
          <derivation type=original>
```

```
            <domain type=art>
            <factuality type=fiction>
            <interaction type=none>
            <preparedness type=prepared>
            <purpose type=entertain degree=high>
            <purpose type=inform degree=medium>
</textDesc>
```

รูป 10.12 ตัวอย่าง <textDesc> ของนวนิยาย

2. <particDesc> เป็นแท็กที่ใช้บรรยายถึงผู้พูด และผู้มีส่วนร่วมในการพูดคุยนั้น ภายในประกอบด้วยแท็ก <person> หรือ <personGrp> มากกว่าหนึ่งแท็กได้ และอาจตามด้วยแท็ก <particLinks> ด้วยหรือไม่ก็ได้ รายละเอียดของแท็กย่อยต่างๆ มีดังนี้

- <person> : ใช้บอกลักษณะของบุคคลผู้ร่วมในการสนทนา คุณสมบัติที่กำหนดได้มีดังนี้

role : ระบุบทบาทของของบุคคลนั้นในกลุ่ม

sex : ระบุเพศของบุคคลนั้น ค่าที่กำหนดได้มี

- m : เพศชาย
- f : เพศหญิง
- u : ไม่ทราบเพศ

age : ระบุกลุ่มอายุของบุคคลนั้น

ตัวอย่างในรูปที่ 10.13 เป็นการบรรยายถึงบุคคลแบบไม่เป็นทางการโดยใช้แท็ก <person>

```
     <person id=P1 sex=F age='mid'>
       <p>Female informant, well-educated, born in
Shropshire UK, 12 Jan 1950, of unknown occupation.
Speaks French fluently. Socio-Economic status B2 in the
PEP classification scheme.
     </person>
```

รูป 10.13 ตัวอย่างการบรรยายบุคคลอย่างไม่เป็นทางการ

- <personGrp> : ใช้บอกลักษณะของกลุ่มหรือคนหลายคนที่ถูกมองเป็นเสมือนบุคคลคนเดียว คุณสมบัติที่กำหนดได้มี

role : ระบุบทบาทที่ปรากฏในการสนทนานั้น

sex : ระบุเพศของกลุ่มนั้น ค่าที่กำหนดได้ ได้แก่

- m : เพศชาย
- f : เพศหญิง
- u : ไม่ทราบเพศ
- x : มีทั้งสองเพศ

age : ระบุกลุ่มอายุ

size : ระบุขนาดของกลุ่ม

- <particLinks> : ใช้บอกลักษณะความสัมพันธ์ทางสังคม ดังในตัวอย่างข้างล่างนี้ ซึ่งอธิบายความสัมพันธ์ของคนเจ็ดคน คือ P1 ถึง P7 โดยที่ P1 และ P2 เป็นพ่อแม่ผู้ปกครองของ P3 P4 ส่วน P1 และ P2 เป็นสามีภรรยากัน และ P1 เป็นนายจ้างของ P3 P5 P6 และ P7

```
<particLinks>
  <relation desc='parent' active='P1 P2'
        passive='P3 P4' mutual=N>
  <relation desc='spouse' active='P1 P2'>
  <relation class='social' desc='employer'
          active=P1 passive='P3 P5 P6 P7' mutual=N>
</particLinks>
```

รูป 10.14 ตัวอย่างการใช้ <particLinks>

3. <settingDesc> เป็นแท็กที่ใช้กำกับข้อมูลเกี่ยวกับสภาพแวดล้อมที่ตัวบทนั้นเกิดขึ้น โดยบอกแต่ละลักษณะภายใต้แท็ก <setting>

- <setting> : ระบุลักษณะสภาพแวดล้อมอย่างหนึ่ง คุณสมบัติที่กำหนดใช้ได้คือ

who : ระบุว่าผู้ร่วมในการสนทนาคือใคร ภายในแท็กนี้อาจประกอบด้วยแท็กย่อยต่างๆ ดังนี้

- <name> : ใช้บอกชื่อเฉพาะ คุณสมบัติที่กำหนดได้คือ

  type : ระบุชนิดของสิ่งที่เรียกชื่อ

- <date> : ใช้บอกวันที่ในรูปแบบต่างๆ คุณสมบัติที่กำหนดคือ

  value : บอกวันเดือนปีในรูปแบบมาตรฐาน เช่น ปี-เดือน-วัน

- <time> : ใช้บอกเวลาในรูปแบบต่างๆ คุณสมบัติที่กำหนด คือ

  value : บอกเวลาตามรูปแบบมาตรฐาน

- <locale> : ใช้บอกสภาพของสถานที่ เช่น ห้อง ภัตตาคาร สวนสาธารณะ เป็นต้น

- <activity> : ใช้บอกถึงการกระทำอื่นๆ ที่กระทำอยู่ระหว่างการพูดของผู้สนทนา

ตัวอย่างในรูปที่ 10.15 และ 10.16 แสดงการกำกับข้อมูลสภาพแวดล้อมโดยวิธีการบอกเล่าแบบง่ายๆ และวิธีการแบบเป็นทางการ ตามลำดับ

```
<settingDesc>
 <p>The time is early spring, 1989.  P1 and P2 are
 playing on the rug of a suburban home in Bedford.
 P3 is doing the washing up at the sink.  P4 (a
 radio announcer) is in a broadcasting studio in London.
</settingDesc>
```

รูป 10.15 ตัวอย่าง <settingDesc> แบบไม่เป็นทางการ

```
<settingDesc>
<setting who="P1 P2">
 <name type=city>Bedford</>
 <name type=region>UK: South East</>
 <date value=1989>early spring, 1989</>
 <locale>rug of a suburban home</>
 <activity>playing</>
</setting>
```

```
 <setting who=P3>
  <name type=city>Bedford</>
  <name type=region>UK: South East</>
  <date value=1989>early spring, 1989</>
  <locale>at the sink</>
  <activity>washing-up</>
 </setting>
 <setting who=P4>
  <name type=place>London, UK</>
  <time>unknown</>
  <locale>broadcasting studio</>
  <activity>radio performance</>
 </setting>
</settingDesc>
```

รูป 10.16 ตัวอย่าง <settingDesc> แบบเป็นทางการ

5.4.2 การกำกับข้อมูลทางภาษาศาสตร์

ในรายงานของ TEI บทที่ 15 ได้กล่าวถึงการกำกับข้อมูลทางภาษาศาสตร์ไว้ ผู้ที่ต้องการสร้างคลังข้อมูลภาษาที่มีการกำกับข้อมูลทางภาษาศาสตร์ต่างๆเพิ่มเติมจึงจำเป็นต้องอ่านบทที่ 15 นี้ด้วย ตัวอย่างของรายการแท็กที่กำหนดไว้ เช่น

- <s> ใช้กำกับขอบเขตประโยค
- <cl> ใช้กำกับอนุพากย์ (clause)
- <phr> ใช้กำกับโครงสร้างวลี
- <w> ใช้กำกับคำ ซึ่งสามารถกำหนดคุณสมบัติ lemma ได้
- <m> ใช้กำกับหน่วยคำ (morpheme) คุณสมบัติที่ใช้คือ baseform ใช้บอกรูปพื้นฐานของหน่วยคำนั้น
- <c> ใช้กำกับตัวอักษร

แท็กทั้งหมดนี้สามารถกำหนดคุณสมบัติสองอย่างได้เหมือนกัน คือ type และ function ส่วนข้อมูลการวิเคราะห์ต่างๆนั้นสามารถใส่ไว้ในคุณสมบัติ ana ได้ เช่น

```
<s type=sentence>
     <w ana=AT >The</w>
     <w ana=NN1>victim</w>
     <m ana=GEN>'s</m>
     <w ana=NN2>friends</w>
     <w ana=VVD>told</w>
     <w ana=NN2>police</w>
     <w ana=CST>that</w>
     <w ana=NP1>Krueger</w>
     <w ana=VVD>drove</w>
     <w ana=II >into</w>
     <w ana=AT >the</w>
     <w ana=NN1>quarry</w>
     <w ana=CC >and</w>
     <w ana=RR >never</w>
     <w ana=VVD>surfaced</w>
     <c ana=PUN>.</c>
</s>
```

รูป 10.17 ตัวอย่างการกำกับข้อมูลหมวดคำ

อีกวิธีหนึ่งคือการใช้แท็ก เพื่อกำหนดขอบเขตของการวิเคราะห์นั้นแล้วใส่ผลการวิเคราะห์ในคุณสมบัติที่ชื่อ value ของ ซึ่งภายในแท็ก จะกำหนดขอบเขตโดยใช้คุณสมบัติ from และ to เพื่อบอกจุดเริ่มต้นและจุดสิ้นสุดของขอบเขตนั้น ดังตัวอย่างในรูป 10.18 ที่ the moment ที่ใช้ใน <s id=MQp1s2p114s5> อ้างถึงข้อความตั้งแต่ <s id=MQp1s2p114s3> ถึง <s id=MQp1s2p114s5>

```
     <p id=MQp1s2p114>
     <s id=MQp1s2p114s1>There was certainly a definite
point at which the thing began.
     <s id=MQp1s2p114s2>It was not; then it was suddenly
inescapable, and nothing could have frightened it away.
</s>
```

```
    <s id=MQp1s2p114s3>There was a slow integration,
during which she, and the little animals, and the moving
grasses, and the sun-warmed trees, and the slopes of
shivering silvery mealies, and the great dome of blue
light overhead, and the stones of earth under her feet,
became one, shuddering together in a dissolution of
dancing atoms.</s>
    <s id=MQp1s2p114s4>She felt the rivers under the
ground forcing themselves pain&shy;fully along her
veins,  swelling them out in an unbearable pressure; her
flesh          was the earth, and suffered growth like a
ferment; and  her eyes stared, fixed like the eye of the
sun.</s>
    <s id=MQp1s2p114s5>Not for one second longer (if the
terms for time apply) could she have borne it; but then,
with a sudden movement forwards and out, the whole
process stopped; and <emph rend=italic>that</emph> was
<soCalled rend=dquo>the moment</soCalled> which it was
impossible to remember afterwards.</s>
    <span resp=DTL
          value='the moment'
          from=MQp1s2p114s3
          to=MQp1s2p114s5>
    <s id=MQp1s2p114s6>For during that space of time
(which    was timeless) she understood quite finally her
smallness, the unimportance of humanity.</s>
```

รูป 10.18 ตัวอย่างการใช้ ในการวิเคราะห์

6. สรุป

ในบทนี้ ได้กล่าวถึงการกำกับข้อมูลภาษาและมาตรฐานที่ใช้ในกำกับข้อมูลซึ่งคือ SGML และ XML พร้อมทั้งอธิบายถึงบทบาทและหน้าที่ของ TEI ที่ได้กำหนด DTD พื้นฐานสำหรับเอกสารประเภทต่างๆ อย่างไรก็ตาม ข้อมูลในบทนี้เป็นเพียงการแนะนำประเด็นหัวข้อต่างๆ เพื่อให้ผู้อ่านสามารถอ่านและทำความเข้าใจเอกสารหรือคลังข้อมูลภาษาที่กำกับตามมาตรฐานของ SGML/XML แต่สำหรับผู้ที่ต้องการสร้างคลังข้อมูลภาษาโดยมีการกำกับข้อมูลตามมาตรฐานของ SGML/XML ยังจำเป็นต้องศึกษาเพิ่มเติมจากรายงานของ TEI โดยละเอียด

พร้อมทั้งศึกษาตัวอย่างของภาษาที่ใช้กำกับข้อมูล (markup language) ที่มีใช้ในโครงการต่างๆ เพื่อเป็นแนวทางในการกำหนดภาษาสำหรับกำกับข้อมูลของตนเอง ให้เป็นไปตามมาตรฐานสากล

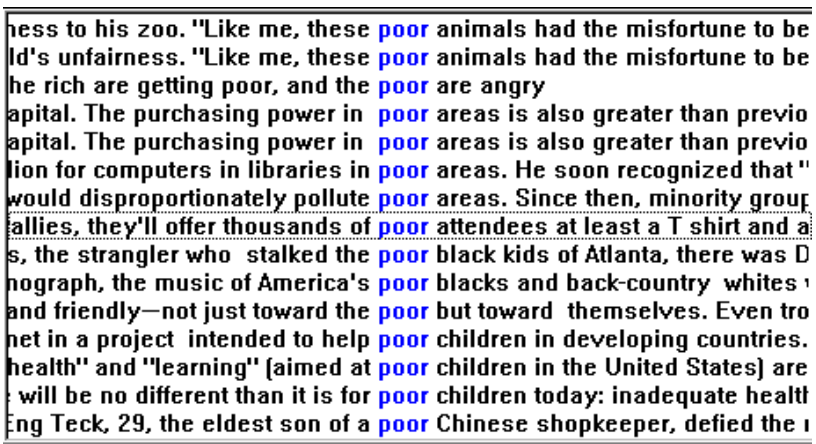
บทที่ 11 โปรแกรมคอนคอร์แด๊นซ์

1. บทนำ

โปรแกรมที่ใช้กับคลังข้อมูลภาษากันอย่างแพร่หลายและเป็นโปรแกรมที่เก่าแก่ที่สุดในการใช้งานด้านมนุษยศาสตร์ คือ โปรแกรมที่เรียกว่าคอนคอร์แด๊นซ์ (concordance) (Oakes 1998:150) หน้าที่หลักของโปรแกรมนี้คือค้นหาคำที่ต้องการและจัดเรียงคำนั้นพร้อมบริบทที่ปรากฏ ทั้งนี้คำนิยามของ concordance คือ "An alphabetical arrangement of the principal words contained in a book, with citations of the passages in which they occur" (OED s.v. 6.b.) ผลที่ได้จากการจัดเรียงจะเป็นรายการของคำที่ต้องการค้นพร้อมทั้งบริบทการเกิดของคำนั้นๆ โดยที่คำที่ต้องการนั้นมักจะจัดเรียงไว้เป็นแนวตรงกันอยู่กลางหน้า ดังตัวอย่างในรูปที่ 11.1

การจัดทำรายการคอนคอร์แด๊นซ์ (concordance) นั้นมีการทำกันมาก่อนที่จะมีคอมพิวเตอร์ใช้ โดยในสมัยนั้นเป็นการทำด้วยแรงงานคนทั้งสิ้น ซึ่งมีการทำกันมานาน เช่น การทำรายการคอนคอร์แด๊นซ์ของคัมภีร์ไบเบิลในปีค.ศ. 1550 โดย John Marbeck การทำรายการคอนคอร์แด๊นซ์บทกวีของ Shakespeare (A Concordance to Shakespeare's Poems) โดย Helen K. Furness ในปีค.ศ.1874 การทำรายการคอนคอร์แด๊นซ์บทกวีของ Matthew Arnold ในปีค.ศ. 1959 ซึ่งจัดทำขึ้นโดย Parrish ผลที่ได้จากการจัดทำรายการคอนคอร์แด๊นซ์นี้สามารถนำมาพิมพ์เป็นหนังสือจำหน่ายแก่ผู้ต้องการศึกษาการใช้ภาษาในงานนั้นโดยละเอียดได้ แต่ในปัจจุบัน เราสามารถใช้โปรแกรมคอมพิวเตอร์ช่วยในการค้นและจัดเรียงได้โดยสะดวก จึงไม่มีความจำเป็นที่จะ

พิมพ์ผลที่ได้ออกมานี้เป็นหนังสือแบบเดิมอีก เพราะว่าผลที่ได้จะมีขนาดใหญ่ เช่น ถ้าจัดทำรายการคอนคอร์แด๊นซ์ของคำทุกคำในงานเขียนของ Chaucer ผลที่ได้จะมีขนาดเท่ากับหนังสือเล่มใหญ่ๆ 10 เล่ม และถ้าพิมพ์ออกเป็นหนังสือจริงก็จะมีราคาแพงมาก อีกทั้งการใช้งานก็จะยุ่งยาก ตรงกันข้ามกับการสืบค้นโดยตรงในคอมพิวเตอร์ เพราะเราสามารถเลือกค้นและจัดเรียงเฉพาะรายการคำที่ต้องการได้ จึงสะดวกมากกว่า และอีกประการหนึ่ง ถ้าหากเราพิมพ์รายการคอนคอร์แด๊นซ์ ออกมาเป็นหนังสือ ผลที่ได้นี้ก็จะมีลักษณะที่ตายตัวแน่นอน เปลี่ยนแปลงไม่ได้ ในขณะที่ถ้าใช้บนคอมพิวเตอร์ เราสามารถปรับการแสดงผลในรูปแบบต่างๆเพื่อให้เข้ากับความต้องการได้มากกว่า เช่น สั่งให้เรียงตามคำที่ปรากฏทางด้านซ้ายของคำหลัก (key word) หรือสั่งให้เรียงตามคำที่ปรากฏทางขวาของคำหลัก เป็นต้น



รูป 11.1 ตัวอย่างรายการคอนคอร์แด๊นซ์ของคำว่า poor

หลักการโดยคร่าวๆในการสร้างรายการคอนคอร์แด๊นซ์ คือการสร้างดัชนีคำ (word index) จากตัวเอกสารก่อน ดัชนีคำอาจเป็นเพียงรายการคำที่ปรากฏในเอกสารนั้นๆ พร้อมทั้งระบุตำแหน่งที่พบคำนั้นๆ ในเอกสาร บางครั้งก็

เรียกดัชนีคำว่าแฟ้มข้อมูลชี้กลับ (converted file) แต่ถ้าแสดงออกมาในรูปของคำพร้อมกับบริบทที่เกิดก็จะเรียกว่าเป็นการแสดงผลแบบแสดงคำหลักพร้อมบริบท (Key Words in Context) ในกรณีที่ไม่ต้องการให้แสดงรายการของคำทุกคำ เช่น คำที่เป็นคำไวยากรณ์ทั้งหลายซึ่งมักมีความถี่สูงมาก ก็สามารถกันคำเหล่านี้ออกไปโดยสร้างรายการคำยกเว้น (stop-list) ได้

2. ประเภทของโปรแกรมคอนคอร์แด๊นซ์

แต่เดิม โปรแกรมคอนคอร์แด๊นซ์ที่ใช้มีการใช้งานอยู่สองลักษณะ คือ ใช้ในลักษณะที่เป็นคำสั่งชุด (batch) หรือใช้ในลักษณะที่เป็นแบบปฏิสัมพันธ์ (interactive) ในแบบแรกนั้น คำสั่งที่จะให้ประมวลผลเอกสารนั้นจะถูกรวมไว้ด้วยกันเป็นชุดคำสั่ง ตัวโปรแกรมคอนคอร์แด๊นซ์จะประมวลผลเอกสารโดยอาศัยชุดคำสั่งนี้ จากนั้นจึงแสดงผลลัพธ์ที่ได้ออกมา หลังจากที่สั่งประมวลผลแล้วผู้ใช้สามารถมารับผลข้อมูลที่ได้ในภายหลัง ตัวอย่างของโปรแกรมคอนคอร์แด๊นซ์ที่ทำงานในลักษณะนี้ คือ Micro-OCP (Oxford Concordance Program) ส่วนโปรแกรมที่ทำงานแบบมีปฏิสัมพันธ์ (interactive) นั้น โปรแกรมสามารถแสดงผลออกมาทันทีที่สั่งให้ค้นหาคำอะไรสักคำ ลักษณะสำคัญของโปรแกรมคอนคอร์แด๊นซ์แบบนี้คือต้องตอบสนองกลับได้เร็ว และเพื่อที่จะให้ประมวลผลได้เร็ว ตัวเอกสารที่ต้องการประมวลอาจจะต้องมีการแปลง (compile) เป็นฐานข้อมูลโดยเฉพาะเพื่อใช้กับโปรแกรมคอนคอร์แด๊นซ์นั้นๆ เช่นในกรณีของโปรแกรม TACT ข้อมูลที่จะค้นได้จะต้องผ่านการแปลงให้เป็นฐานข้อมูลโดยเฉพาะของโปรแกรม TACT ก่อน แต่ข้อเสียของวิธีนี้ คือ ถ้ามีการแก้ไขเพิ่มเติมข้อมูลในคลังข้อมูลภาษา ก็จะต้องทำการแปลง (compile) ข้อมูลใหม่ทุกครั้ง วิธีนี้จึงเหมาะกับการใช้งานที่ไม่มีการเปลี่ยนแปลงคลังข้อมูลบ่อย

แต่ในปัจจุบัน ความแตกต่างระหว่างการทำงานแบบใช้คำสั่งชุดหรือแบบมีปฏิสัมพันธ์ไม่ใช่เรื่องสำคัญมากนัก เพราะประสิทธิภาพของคอมพิวเตอร์ที่เพิ่มมากขึ้นทุกวันทำให้ไม่มีความจำเป็นใดๆที่จะต้องใช้งานแบบคำสั่งชุด และโปรแกรมที่ทำงานแบบมีปฏิสัมพันธ์ก็อาจไม่จำเป็นต้องแปลงข้อมูลก่อนใช้ แต่สามารถประมวลผลตัวเอกสารได้โดยตรง และสามารถตอบสนองกลับได้รวดเร็วพอโดยเฉพาะอย่างยิ่งเมื่อเอกสารที่ต้องการค้นนั้นไม่ได้มีขนาดใหญ่มากเกินไป ดังนั้นจึงแทบจะไม่มีความจำเป็นใดๆ ที่ต้องใช้วิธีการประมวลผลแบบใช้คำสั่งชุด

3. การค้นข้อมูล

ในโปรแกรมคอนคอร์แด๊นซ์ทั่วๆ ไป ผู้ใช้สามารถกำหนดให้ค้นหาข้อมูลที่ต้องการในลักษณะต่างๆ ดังต่อไปนี้

- ให้คำหลัก (keyword) ที่ต้องการจะค้น ซึ่งอาจค้นหาคำได้มากกว่าหนึ่งคำ โดยใช้ตัวเชื่อมทางตรรกกะ (boolean operator) เช่น AND OR NOT ประกอบด้วยได้
- ให้ค้นโดยกำหนดข้อความที่มีอักขระพิเศษ (wildcard) เช่น ใช้ * แทนตัวอักษรอะไรกี่ตัวก็ได้ ใช้ ? แทนตัวอักษรอะไรก็ได้เพียงหนึ่งตัว เป็นต้น ตัวอย่างเช่น ถ้าให้ค้น *ing ในโปรแกรม AntConc จะหมายถึงให้ค้นหาคำทุกคำที่ลงท้ายด้วย ing หรือบางโปรแกรมสามารถค้นโดยใช้รูปแบบการค้นแบบ regular expression ได้ เช่น โปรแกรม WordSmith โปรแกรม AntConc
- ให้ค้นโดยระบุแท็กที่ต้องการด้วย ในบางโปรแกรม ผู้ใช้สามารถระบุให้ค้นหาข้อมูลที่มีแท็กตามที่กำหนดได้ ถ้าหากว่าเอกสารนั้นถูกกำกับข้อมูลเพิ่มเติมมาแล้ว เช่น อาจให้ค้นเฉพาะข้อความในส่วนที่เป็นชื่อเรื่อง (title) เป็นต้น

4. การแสดงผล

โปรแกรมคอนคอร์แด๊นซ์นั้นสามารถแสดงผลที่ได้ออกมาในหลายลักษณะ บางโปรแกรมก็สามารถเลือกแสดงผลได้มากกว่าหนึ่งแบบ

- การแสดงผลแบบแสดงคำหลักพร้อมบริบท KWIC (Key-Word In Context) เป็นการแสดงผลแบบที่ยึดคำค้นเป็นหลักอยู่ตรงแกนกลางพร้อมกับแสดงบริบททั้งซ้ายและขวาของคำนั้น ซึ่งอาจแสดงผลโดยยึดขอบเขตของประโยคเป็นหลัก หรือยึดตามจำนวนคำหรือตัวอักษรที่ปรากฏข้างซ้ายและขวาของคำหลักนั้น หรืออาจกำหนดโดยยึดจำนวนบรรทัดก่อนและหลังของคำค้นนั้นก็ได้ หากต้องการแสดงตามจำนวนบรรทัด De Tollenaere (1973) ได้แนะว่าควรแสดง 4 บรรทัดก่อนคำนั้นและอีก 2 บรรทัดหลังคำนั้น เพื่อที่จะได้บริบทเพียงพอสำหรับงานด้านพจนานุกรม (lexicography)
- การแสดงผลแบบเต็มจอ (Full Text) เป็นการแสดงผลในลักษณะที่แสดงเอกสารนั้นตามที่ปรากฏจริงๆ เต็มหน้าจอ โดยที่คำที่ต้องการค้นจะถูกเน้นให้เห็นเด่นชัดกว่าคำอื่นๆ เช่น ทำเป็นตัวหนา หรือขีดเส้นใต้ เป็นต้น

นอกจากการแสดงผลแบบ KWIC หรือแบบเต็มจอแล้ว โปรแกรมคอนคอร์แด๊นซ์โดยทั่วไปยังสามารถแสดงผลลัพธ์อื่นๆ นอกเหนือจากการจัดเรียงคำได้ด้วย เช่น

- แสดงรายการความถี่ของคำ (word frequency list) คือ แสดงรายการคำทั้งหมดที่ปรากฏในข้อมูลว่าแต่ละคำมีความถี่ของการใช้เท่าใด
- แสดงค่าสถิติการใช้ตัวอักษรต่างๆ คือ แสดงผลการวิเคราะห์การใช้ตัวอักษรทั้งหมดที่มีปรากฏในข้อมูลว่า ตัวอักษรใดมีการใช้มากน้อยเพียงใด
- แสดงการปรากฏร่วมกันของคำ (collocation) คือ แสดงให้เห็นว่าคำที่ปรากฏร่วมกับคำหลักที่ค้นนั้นมีอะไรบ้าง โดยอาจใช้วิธีง่ายๆ คือแสดงรายคำที่ปรากฏ

ด้านซ้ายหรือขวาของคำหลักนั้นในขอบเขตที่กำหนดเช่น 5 คำข้างซ้ายและ 5 คำข้างขวา หรือบางโปรแกรมก็สามารถคำนวณค่าทางสถิติที่ใช้แสดงระดับนัยสำคัญทางสถิติของการปรากฏร่วมกัน เพื่อดูว่าคำใดบ้างที่น่าสงสัยว่าเป็นคำที่ปรากฏร่วมกับคำหลักที่ต้องการค้น

5. ข้อจำกัดของโปรแกรมคอนคอร์แด๊นซ์

โปรแกรมคอนคอร์แด๊นซ์ไม่สามารถวิเคราะห์ด้านวจีวิภาค (Morphology) ของแต่ละภาษาได้โดยอัตโนมัติ ทั้งนี้เพราะโปรแกรมคอนคอร์แด๊นซ์ทำงานโดยพิจารณาตัวอักษรที่ปรากฏอยู่ในเอกสารนั้นเป็นหลัก เช่น เมื่อค้นคำว่า *write* โปรแกรมอาจจะไม่สนใจคำว่า *writing wrote written* เป็นต้น ดังนั้น ผู้ใช้ต้องกำหนดรูปแบบการค้นให้ครอบคลุมวิภัตติปัจจัยแบบต่างๆ เอง เช่น อาจค้นโดยกำหนดให้ดูรากคำ (stem) เป็นหลัก และให้ค้นเพิ่มเติมโดยกำหนดส่วนหน้าหรือหลังตามอุปสรรค (prefix) หรือปัจจัย (suffix) ที่เป็นไปได้ วิธีนี้ผู้ใช้จะต้องทำการวิเคราะห์ตัวภาษานั้นมาก่อน คือ ต้องสร้างรายการของรูปแบบที่ต้องการจะค้น เพื่อให้ได้ข้อมูลแบบที่ต้องการมากที่สุด เช่น ต้องรู้ว่ารูปแปร (variant form) ของศัพท์ที่ต้องการมีอะไรบ้าง อีกวิธีหนึ่ง คือ ผู้ใช้อาจจะต้องใส่แท็กเพื่อบอกข้อมูลเพิ่มเติมเช่น บอกรากคำของแต่ละคำไปเลย แล้วจึงให้ค้นจากรากคำนั้นเป็นหลัก

ในงานทางด้านวากยสัมพันธ์และอรรถศาสตร์ (syntax and semantics) ปกติโปรแกรมคอนคอร์แด๊นซ์จะไม่สามารถค้นหาข้อมูลพวกนี้ได้ นอกจากจะได้มีการวิเคราะห์ทางด้านวากยสัมพันธ์ และอรรถศาสตร์ไว้ก่อน และกำกับข้อมูลเหล่านั้นในรูปของแท็กต่างๆเพิ่มเติมไว้ในเอกสาร

6. ประโยชน์ของโปรแกรมคอนคอร์แด๊นซ์

โปรแกรมคอนคอร์แด๊นซ์ช่วยให้เราสามารถค้นหาคำที่ต้องการว่ามีบริบทการใช้จริงๆอย่างไรบ้าง โปรแกรมคอนคอร์แด๊นซ์จึงเป็นเครื่องมือพื้นฐานสำหรับการใช้ในงานต่างๆที่เกี่ยวข้องกับการศึกษาภาษา ประโยชน์ของการใช้โปรแกรมคอนคอร์แด๊นซ์จึงเหมือนกับการใช้ประโยชน์ของคลังข้อมูลภาษาที่มีต่อการศึกษาวิจัยภาษา ต่อการสอนภาษา ซึ่งได้กล่าวถึงไปแล้วในบทที่ 2 และ 3 ในที่นี้จึงขอกล่าวสรุปแต่เพียงคร่าวๆ ดังนี้

6.1 ประโยชน์ด้านการวิจัยทางภาษาศาสตร์

Sinclair (1986:202) กล่าวถึงบทบาทของโปรแกรมคอนคอร์แด๊นซ์ว่า "challenge our current linguistic descriptions quite fundamentally", freeing linguists from reliance on intuition and enabling them to "find explanations that fit the evidence, rather than adjusting the evidence to fit a pre-set explanation." กล่าวคือ การใช้โปรแกรมคอนคอร์แด๊นซ์กับคลังข้อมูลภาษาจะช่วยให้นักภาษาศาสตร์ได้เห็นรูปแบบการใช้ภาษาที่เกิดขึ้นจริง รูปแบบที่อาจจะนึกไปไม่ถึง และช่วยให้สามารถหาคำอธิบายภาษาให้สอดคล้องกับความเป็นจริง อันจะเป็นการเปลี่ยนโฉมหน้าของการศึกษาภาษาศาสตร์ได้ ตัวอย่างเช่น Fox (1991) ใช้คลังข้อมูลภาษาจาก COBUILD พิสูจน์ว่าคำกริยาในภาษาอังกฤษนั้นส่วนมากไม่ได้เป็นคำกริยาที่เป็นสกรรมหรืออกรรมอย่างเด็ดขาด (absolute transitive or intransitive) Higgins (1991) ใช้โปรแกรมคอนคอร์แด๊นซ์หารูปแบบการปรากฏ *the *ing of* เพื่อหาสาเหตุว่าทำไมนิสิตคนหนึ่งของเขาถึงได้ใช้วลีว่า *the remaining of his army* Tribble (1989) ใช้โปรแกรมคอนคอร์แด๊นซ์ เพื่อหาความแตกต่างของการใช้คำหน้าที่ (function word) ระหว่าง

เจ้าของภาษา (native speaker) กับผู้ที่ไม่ได้เป็นเจ้าของภาษา (non-native speaker) ซึ่งก็พบว่ามีความแตกต่างของการใช้คำ เช่น *however*, *moreover*, *therefore* อย่างมาก

6.2 ประโยชน์ด้านการสอนภาษา

โปรแกรมคอนคอร์แด๊นซ์ จะแสดงรายการตัวอย่างการใช้จริงของคำหรือวลีที่เกิดขึ้น ผู้เรียนสามารถเรียนรู้จากตัวอย่างภาษาเหล่านั้นด้วยตัวเอง วิธีนี้จะเปลี่ยนแปลงพื้นฐานการเรียนการสอนภาษาจากการฟังอาจารย์บอกเป็นการเรียนด้วยข้อมูลจริงมากขึ้น แต่เหตุผลที่โปรแกรมคอนคอร์แด๊นซ์ ยังไม่เป็นที่แพร่หลายในด้านนี้ อาจเป็นเพราะรูปแบบการเรียนการสอนภาษายังเป็นแบบเก่า คือ อาจารย์เป็นผู้บอกกฎเกณฑ์ต่างๆ คำถามคือ อาจารย์ผู้สอนภาษาจะเปลี่ยนแนวคิดด้านการเรียนการสอนในแบบเดิมได้ไหม จะยอมรับได้ไหมถ้าผู้เรียนใช้โปรแกรมคอนคอร์แด๊นซ์ แล้วพบว่าบางคำหรือบางโครงสร้างไม่ได้มีการใช้ตามที่อาจารย์สอนเสมอไป นอกจากนี้ Louw (1991) เสนอว่าถ้าเราใช้โปรแกรมคอนคอร์แด๊นซ์ แสดงรายการคำที่มีความถี่ของการใช้มาก 2,000 คำแรกในภาษาอังกฤษ และสอนนักเรียนให้เข้าใจการใช้คำเหล่านั้นเป็นอย่างดี ก็จะเป็นการสอนภาษาอังกฤษให้นักเรียนต่างชาติที่มีประสิทธิภาพวิธีหนึ่ง และ Higgins (1991) กล่าวเพิ่มเติมว่าโปรแกรมคอนคอร์แด๊นซ์เป็นโปรแกรมที่มีคุณค่าอย่างมากในการใช้ประโยชน์จากคอมพิวเตอร์เพื่อการเรียนการสอนภาษา เพราะจะทำให้ผู้เรียนได้เห็นข้อมูลมหาศาลของการใช้ภาษาจริงๆที่เกิดขึ้น

6.3 ประโยชน์ในงานประยุกต์อื่น ๆ

โปรแกรมคอนคอร์แด๊นซ์ ยังสามารถนำมาประยุกต์ใช้กับงานอื่นๆ ได้ เช่น Catherine N. Ball (1996) กล่าวถึงการใช้โปรแกรมคอนคอร์แด๊นซ์เพื่อ

ตรวจสอบการแอบคัดลอกงานของผู้อื่น (plagiarism) โดยใช้การเปรียบเทียบตัวบทจากรายงานของนักเรียนและตัวบทจากหนังสือที่สงสัยว่าจะเป็นต้นฉบับ ในตัวอย่างข้างล่าง เป็นการเปรียบเทียบรายงานที่นักเรียนส่งกับรายการคำที่ได้จากหนังสือ "Pragmatics and Natural Language Understanding" ของ Green (1989) ซึ่งสามารถทำได้โดยรวมตัวบทจากหนังสือเล่มนี้กับรายงานของนักเรียนเป็นแฟ้มข้อมูลเดียวกันและใส่ตัวอักษร G หรือ S เพื่อบอกที่มาของตัวบทนั้น โดยที่ G แทนตัวบทจากหนังสือของ Green และ S แทนตัวบทจากรายงานนักเรียน จากนั้นใช้โปรแกรมคอนคอร์แด็นซ์เรียงคำในเอกสารโดยกำหนดให้บริบททางซ้ายเป็น 0 และบริบททางขวาเป็น 40 คำ เมื่อได้รายการของคำทั้งหมดแล้ว จึงสังเกตรายการที่มีความคล้ายคลึงกันมาก จากบรรทัดที่ mark ด้วย G และ S เพราะอาจเป็นกรณีของการแอบคัดลอกงานของผู้อื่นได้ ซึ่งเมื่อพบข้อความที่คล้ายคลึงกัน ก็สามารถตรวจดูในตัวบททั้งสองตามบรรทัดที่แสดงไว้ในรายการคอนคอร์แด็นซ์ได้ รูปที่ 11.2 แสดงผลที่ได้จากโปรแกรมคอนคอร์แด็นซ์ และรูปที่ 11.3 เปรียบเทียบตัวบท ณ ตำแหน่งที่สงสัยว่ามีการคัดลอกงาน

```
A 19
G 11 A, and the extent to which the act
G 13 A are members. On the basis of this
G 3 a commodity they call face. FACE is
G 18 a desire to avoid interfering with A's
S 43 a desire to avoid interfering with the
S 29 a face-threatening act FTA are: 1 The
G 9 a face-threatening act in the course of
S 27 a face-threatening act in the course of
G 16 A feel good, and feel that her values are
G 14 a forgo trading in face and perform the
G 16 a NEGATIVE POLITENESS strategy of
S 42 a Negative Politeness strategy, that is,
S 25 a normal conversation is to risk losing
G 15 a POSITIVE POLITENESS strategy of making
S 40 a Positive Politeness strategy, that is,
G 1 a slightly different perspective on
S 29 a speaker's calculation of how much she
```

```
G 20 A the option of not acknowledging
G 11 A, the relative social power of S and A,

about 1
S 38 about trading face and perform the act

accomplishing 2
G 9 accomplishing their goals. According to
S 27 accomplishing their goals. The factors
```

รูป 11.2 ตัวอย่างรายการคอนคอร์แด๊นซ์ที่สร้างจากรายงานนักเรียนและต้นฉบับ

7. ข้อควรระวัง

แม้การใช้โปรแกรมคอนคอร์แด๊นซ์จะเป็นประโยชน์ต่อการใช้คลังข้อมูลภาษา แต่โปรแกรมคอนคอร์แด๊นซ์นี้ก็มีข้อจำกัดบางประการ Catherine Ball (1996) จึงเสนอว่าควรพิจารณาถึงข้อจำกัดเหล่านี้ด้วย ในการใช้โปรแกรมนี้

1. โปรแกรมคอนคอร์แด๊นซ์จะค้นหาโดยยึดรูปคำที่ปรากฏเป็นหลัก ผลที่ได้ออกมาอาจไม่ใช่สิ่งที่ผู้วิจัยต้องการทั้งหมด เช่น ในกรณีของคำพ้องรูป ผู้วิจัยจะได้คำในความหมายที่ไม่ต้องการติดมาด้วย และรายการคำที่ผู้วิจัยต้องการบางครั้งอาจจะถูกละเลยไป เช่น คำๆ เดียวที่มีรูปแปรได้มากกว่าหนึ่งแบบ ดังนั้นผู้วิจัยจึงต้องกำหนดรูปแบบของการค้นหาให้ถูกต้องและครบถ้วนให้มากที่สุด

Student paper	Source
	Green (1989:144)
Face is defined as consisting of the freedom to act unimpeded (negative face) and the satisfaction of having one's values approved of (positive face) (Brown & Levinson, 1987:67).	face is defined as consisting of the freedom to act unimpeded (negative face) and the satisfaction of having one's values approved of (positive face) (Brown & Levinson, 1978, p. 67).
	Green (1989:144)
According to them, to engage in a normal conversation is to risk losing face. Maintaining face requires the cooperation of others, thus interactants trade in this commodity, paying face whenever they perform a face-threatening act in the course of accomplishing their goals.	To engage in normal interaction is to risk losing face ... Maintaining face requires the cooperation of others ..., so interactants trade in face, paying face whenever they must perform a face-threatening act in the course of accomplishing their goals.

รูป 11.3 เปรียบเทียบตัวบทจากรายงานนิสิตและต้นฉบับ

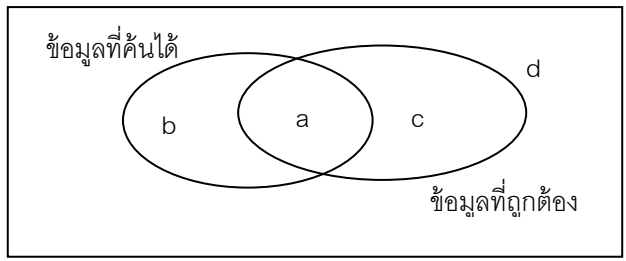
2. โปรแกรมคอนคอร์แด๊นซ์ไม่สามารถใช้ค้นหาคำลดรูปหรือสูญรูป (deletion หรือ zero) ได้เนื่องจากเป็นสิ่งที่ไม่มีปรากฏในเอกสาร ถ้าจำเป็นต้องศึกษาปรากฏการณ์ภาษาเหล่านี้ ผู้วิจัยจำเป็นต้องกำกับข้อมูลเพิ่มเติมลงในเอกสารเพื่อระบุตำแหน่งและคุณสมบัติของสิ่งที่ละไป

3. ต้องระวังอย่าเลือกใช้ค่าความถี่ที่ได้จากโปรแกรมคอนคอร์แด๊นซ์เพียงเพราะเป็นตัวเลขที่ได้มาโดยง่าย ผู้ใช้ควรจะต้องคำนึงถึงรายละเอียดของงานวิจัยนั้นๆ ด้วย ตัวอย่างเช่น ในกรณีของผู้ที่ต้องการศึกษาเปรียบเทียบการใช้คำคุณศัพท์เพศหญิง (feminine) ระหว่างนักเขียนชายและนักเขียนหญิง ซึ่งผู้วิจัยสามารถทำโดยแสดงรายคำคุณศัพท์เหล่านั้นพร้อมทั้งหาความถี่ของคำเหล่านั้น จากนั้นจึงคิดเป็นร้อยละของคำคุณศัพท์ที่ได้เปรียบกับจำนวนคำทั้งหมด ทั้งนี้เพราะตัวเลขจำนวนคำทั้งหมดเป็นตัวเลขที่โปรแกรมคอนคอร์แด๊นซ์สามารถระบุได้ทันที แต่อันที่จริงแล้ว ผู้วิจัยควรคำนวณร้อยละโดยเทียบกับจำนวนคำคุณศัพท์ทั้งหมดที่พบในเอกสารมากกว่าที่จะเทียบกับจำนวนคำทั้งหมด ซึ่งตัวเลขจำนวนคำคุณศัพท์ทั้งหมดนี้โปรแกรมคอนคอร์แด๊นซ์ไม่สามารถระบุได้โดยตรง เนื่องจากโปรแกรมจะไม่สามารถนับจำนวนคำคุณศัพท์ได้ (ยกเว้นในกรณีที่เอกสารนั้นมีการกำกับหมวดคำไว้ด้วย)

4. สิ่งที่ผู้วิจัยต้องการหาอาจไม่สามารถหาได้จากคลังข้อมูลภาษาที่มีอยู่ เช่น งานวิจัยบางเรื่องอาจจำเป็นต้องใช้ข้อมูลภาษาที่ยังไม่มีการพิมพ์เก็บไว้ในคอมพิวเตอร์ ผู้วิจัยจึงไม่ควรเอาข้ออ้างของการที่สามารถใช้โปรแกรมคอนคอร์แด๊นซ์ได้สะดวกมาเพื่อละเลยข้อมูลอื่นๆที่ยังไม่ได้อยู่ในคอมพิวเตอร์ กล่าวคือหากจำเป็นต้องใช้ข้อมูลเหล่านั้นกับโปรแกรมคอนคอร์แด๊นซ์จริงๆ ผู้วิจัยก็ควรจะพิมพ์ข้อมูลเหล่านั้นลงในคอมพิวเตอร์ด้วย

5. ผู้วิจัยควรคำนึงถึงเรื่องความแม่นยำ (precision) และความครบถ้วน (recall) ด้วย โดยที่ค่าความแม่นยำ (precision) ค่าความครบถ้วน (recall) และค่าความผิดพลาด (fallout) เป็นเรื่องทางสถิติที่ใช้แสดงเปรียบเทียบสิ่งที่สังเกตุได้กับสิ่งที่เป็นจริง รูปที่ 11.4 ข้างล่างแสดงเปรียบเทียบการคำนวณค่าทั้งสาม

โดยที่ในการค้นหาข้อมูลใดใด ส่วนของ a+b คือสิ่งที่ค้นมาได้ ซึ่งอาจไม่ใช่สิ่งที่ถูกต้องหรือสิ่งที่ต้องการทั้งหมด ถ้าสิ่งที่ถูกต้องคือ a+c การค้นครั้งนี้ก็จะให้ค่าความแม่นยำจะเท่ากับ a / (a+b) คือมีเฉพาะส่วน a ของสิ่งที่ค้นมาเท่านั้นที่ถูกต้อง ส่วนค่าความครบถ้วนจะเท่ากับ a / (a+c) เพราะเป็นการเทียบกับข้อมูลที่ถูกต้องทั้งหมด (คือ a รวมกับ c) ส่วนค่าความผิดพลาดนั้น คือจากสิ่งที่ค้นมาแล้วไม่ใช่ข้อมูลที่ต้องการเทียบกับข้อมูลทั้งหมดที่ไม่ต้องการ จึงมีค่าเท่ากับ b / (b+d)




```
precision     = a / (a+b)          (accuracy)
recall        = a / (a+c)          (true positive)
fallout       = b / (b+d)          (false positive)
```

รูป 11.4 แสดงการคำนวณค่าความแม่นยำและความครบถ้วน

ในเรื่องของการใช้โปรแกรมคอนคอร์แด๊นซ์นี้ a คือ จำนวนหน่วยที่โปรแกรมหาได้และเป็นหน่วยที่ต้องการ b คือจำนวนหน่วยที่โปรแกรมหาได้แต่ไม่ใช่หน่วยที่ต้องการ c คือ จำนวนหน่วยที่โปรแกรมค้นหาไม่พบแต่เป็นหน่วยที่ผู้วิจัยต้องการ d คือ จำนวนหน่วยที่โปรแกรมค้นหาไม่พบและก็ไม่ใช่หน่วยที่

ต้องการ (โดยทั่วไปผู้วิจัยมักจะไม่ใช้ค่าความผิดพลาด เนื่องจากค่า d มักจะเป็นค่าที่คำนวณไม่ได้หรือหาได้ยาก)

ดังนั้น ค่าความแม่นยำ (precision) จะคำนวณออกมาเป็นร้อยละของจำนวนหน่วยที่ค้นได้ตรงตามความต้องการ ค่าความครบถ้วน (recall) เป็นร้อยละของจำนวนหน่วยที่เกี่ยวข้องที่โปรแกรมค้นหาได้ โดยทั่วไป ผู้วิจัยมักต้องการจะได้ค่าความแม่นยำ 100% และค่าความครบถ้วน 100% แต่ในความเป็นจริงเป็นเรื่องที่เป็นไปได้ยาก ถ้าผู้วิจัยต้องการให้ได้ความครบถ้วนสมบูรณ์ ผู้วิจัยก็มักจะได้สิ่งอื่นๆที่ไม่ใช่สิ่งที่ต้องการมาด้วยมาก ส่งผลให้ได้ค่าความแม่นยำต่ำ (poor precision) แต่ถ้าผู้วิจัยต้องการให้ได้ค่าความแม่นยำสมบูรณ์ โอกาสที่จะพลาดรายการหลายๆรายการก็มีสูง ส่งผลให้ได้ค่าความครบถ้วนต่ำ (poor recall)

ดังนั้น ในการใช้คลังข้อมูลภาษา ผู้วิจัยจึงควรพิจารณาเรื่องพวกนี้ด้วย หากสามารถระบุถึงตัวเลขเหล่านี้ได้ก็ควรจะทำ แต่ในกรณีที่คลังข้อมูลภาษามีขนาดใหญ่มากก็อาจเป็นไปได้ยากที่จะหาหน่วยที่โปรแกรมคอนคอร์แด๊นซ์สืบค้นไม่พบ (missing token) ในคลังข้อมูลภาษานั้น เนื่องจากผู้วิจัยจะต้องหาด้วยตนเองว่าโปรแกรมสืบค้นไม่พบเป็นจำนวนเท่าใด แต่ก็อาจทำได้โดยวิธีการสุ่มตัวอย่างข้อมูลเพื่อประมาณค่าความแม่นยำและความครบถ้วนที่พบ

8. ตัวอย่างการใช้โปรแกรมคอนคอร์แด๊นซ์

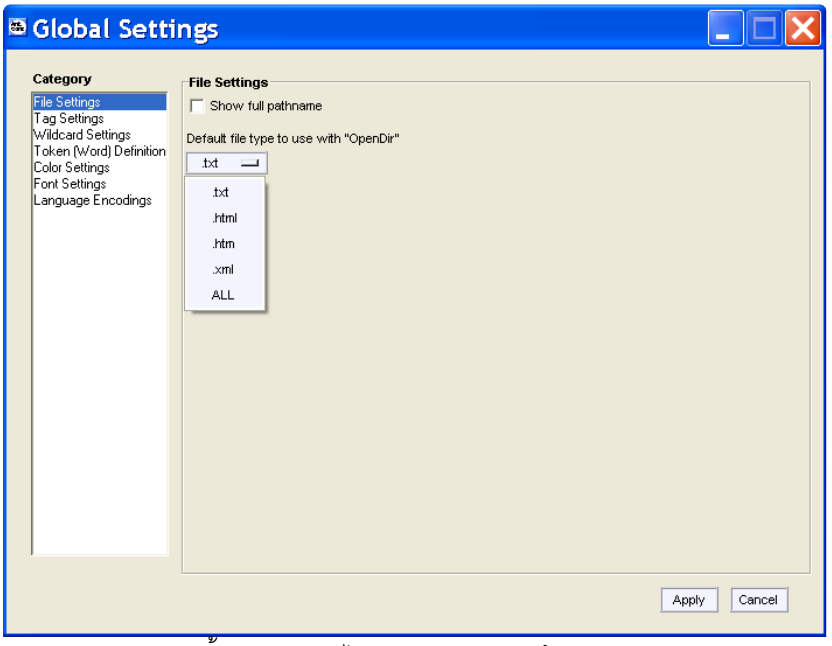
ในตอนนี้ จะเป็นการยกตัวอย่างการใช้โปรแกรมคอนคอร์แด๊นซ์ ในที่นี้เลือกโปรแกรม AntConc เป็นตัวอย่างการใช้งาน เนื่องจากเป็นโปรแกรมที่ใช้งานได้ฟรี สามารถดาว์นโหลดได้จากเว็บ http://www.antlab.sci.waseda.ac.jp/

antconc_index.html เวอร์ชั่น ณ เดือนมีนาคม 2007 คือ AntConc 3.2.1 โปรแกรมที่ดาว์นโหลดมาแล้วสามารถเรียกใช้งานได้ทันที ไม่จำเป็นต้องติดตั้ง

ในการทดลองใช้โปรแกรมนี้ จะใช้ข้อมูลงานเขียนของ Jane Austen จำนวน 9 เล่มซึ่งสามารถดาว์นโหลดหนังสือเหล่านี้ได้จาก Oxford Text Archive (http://www.ota.ox.ac.uk/) และงานเขียนทั้งหมดของ William Shakespeare ซึ่งสามารถดาว์นโหลดได้จาก http://www.it.usyd.edu.au/~matty/Shakespeare/ คลังข้อมูลทดสอบนี้จึงเป็นคลังข้อมูลที่แทนภาษานิยายของนักเขียนทั้งสอง

8.1 การโหลดคลังข้อมูลเข้าโปรแกรม

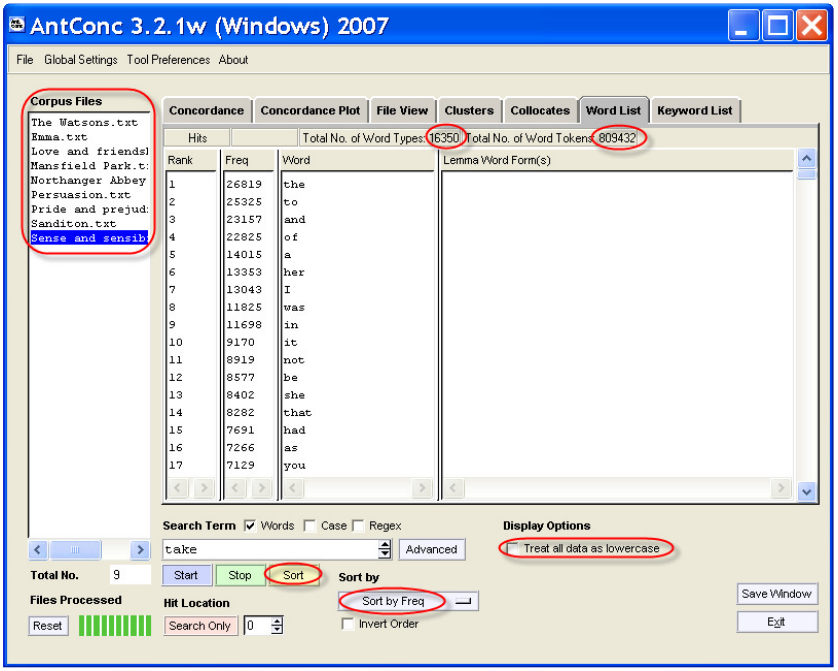
เริ่มต้นด้วยการบอกโปรแกรมให้โหลดคลังข้อมูลโดยเลือก File – Open File(s) แล้วเลือกไฟล์หนังสือทั้ง 9 เล่ม ชื่อไฟล์ทั้งหมดจะปรากฏที่คอลัมน์ซ้ายมือ กรณีที่ต้องการเปิดทุกไฟล์ในไดเร็กเตอรี่นั้น ให้เปิดด้วย File – Open Dir แล้วเลือกไดเร็กเตอรี่ที่เก็บคลังข้อมูลที่ต้องการ เช่น กรณีข้อมูลงานของ Shakespeare ที่โหลดมาจะประกอบด้วยไดเร็กเตอรี่ย่อยสี่อันแยกตามประเภทของงานเขียนเป็น comedies, histories, poetry, tragedies เริ่มต้นโปรแกรมจะถือว่าไฟล์ข้อมูลเก็บในสกุล .txt หากต้องการแก้ไขให้เปิดไฟล์สกุลอื่นๆได้ ให้แก้ไขที่ Global Setting แล้วเลือกสกุลไฟล์จาก .txt เป็นแบบที่ต้องการเช่น ALL คือทุกไฟล์ แล้วกดปุ่ม Apply แล้วจึงเปิดไดเร็กเตอรี่ที่ต้องการได้



รูป 11.5 การตั้งค่านามสกุลไฟล์สำหรับการอ่านข้อมูลแบบ Open Dir

8.2 การหารายการความถี่ของคำ

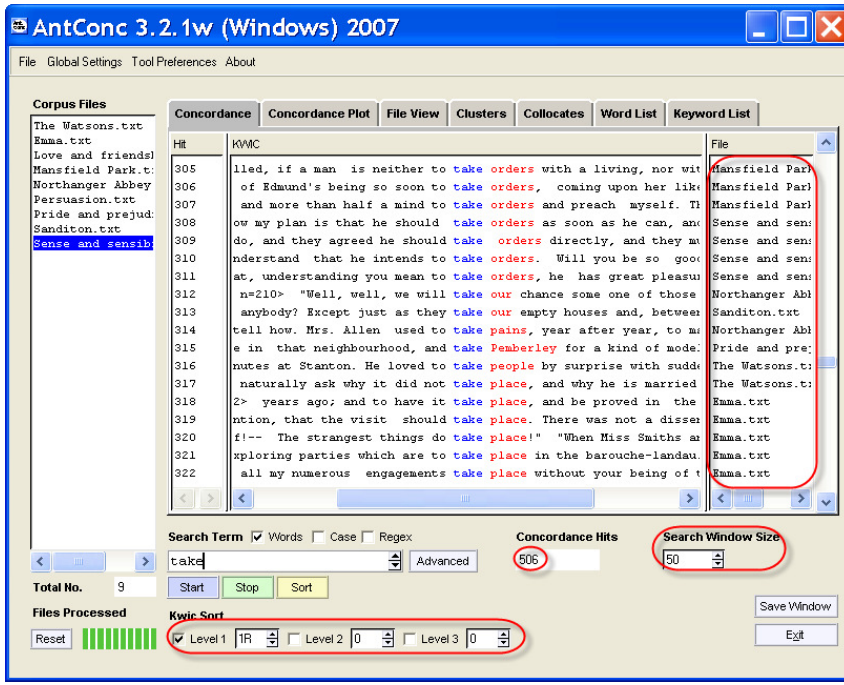
ดูรายการคำทั้งหมด โดยเลือกแท็ป Word List แล้วกดปุ่ม start จะได้รายการคำพร้อมความถี่ปรากฏให้เห็น จำนวนศัพท์ทั้งหมดที่พบในคลังข้อมูลคือตัวเลขของ Word Type ส่วนจำนวนคำทั้งหมดที่พบคือตัวเลขของ Word Token ในการนับคำนั้น เราสามารถเลือกได้ว่าจะแยกความต่างระหว่างตัวพิมพ์ใหญ่หรือเล็กหรือไม่ และรายการคำทั้งหมดนี้ก็สามารถจัดเรียงได้ในหลายลักษณะโดยการเลือก Sort by Freq, Sort by Word, Sort by Word End การใช้ Sort by Word End จะช่วยแสดงคำซึ่งมี suffix เดียวกันมาเป็นกลุ่มให้เห็น



รูป 11.6 หน้าจอ word frequency list

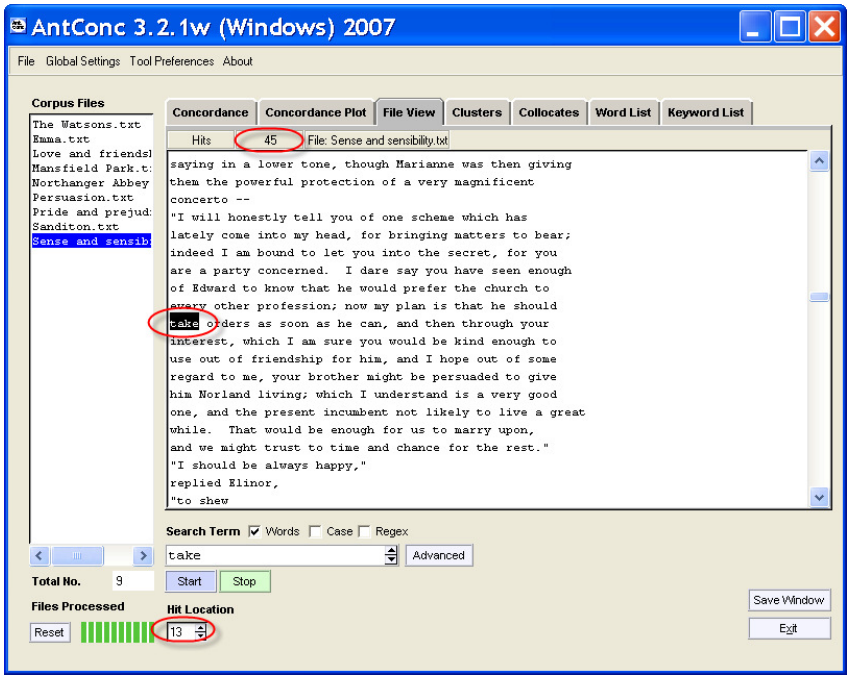
8.3 การค้นคำหลักพร้อมบริบท

การใช้ Concordance ให้เลือกแท็ป Concordance แล้วพิมพ์คำที่ต้องการ เช่น คำว่า take เราสามารถจัดเรียงผลใหม่ได้ โดยเลือกการ sort Level 1, 2, 3 ตามที่ต้องการ 1R หมายถึงคำทางขวาตำแหน่งแรกของคำหลัก จำนวนตัวอย่างที่พบทั้งหมดจะบอกใน Concordance Hits (ในตัวอย่างนี้คือ 506 ตัวอย่าง) ขอบเขตความยาวของข้อมูลที่ต้องการแสดงสามารถกำหนดได้ใน Search Windows size ค่าเริ่มต้นคือ 50 คำ ที่คอลัมน์ขวาที่พบตัวอย่างบรรทัดที่นำมาแสดงจะบอกชื่อไฟล์ข้อมูลที่พบตัวอย่างนั้น



รูป 11.7 หน้าจอ concordance

เราสามารถเลือกคำที่ต้องการมาแสดงในหน้าจอ Concordance ได้ทันทีจากหน้าจอของ Word List โดยการกดที่คำที่ต้องการ เช่น กดที่คำ take ในหน้าจอ Word List ก็จะได้ผลเช่นเดียวกัน (สังเกตดูจะเห็นว่ามีรูปมือชี้ให้กดได้ในหน้าจอ Word List) ในทำนองเดียวกัน ในหน้าจอ Concordance นี้ หากเรากดที่คำหลักในตัวอย่างที่ต้องการ ก็จะนำเราไปสู่หน้าจอ File View โดยแสดงข้อความต่อเนื่องในไฟล์ที่พบตัวอย่างนั้นพร้อมทั้งระบายสีคำหลักที่เลือก ที่ด้านบนจะบอกจำนวนตัวอย่างทั้งหมดที่พบในไฟล์นี้ ที่ด้านล่างแสดงว่าเป็นตัวอย่างที่เท่าไร

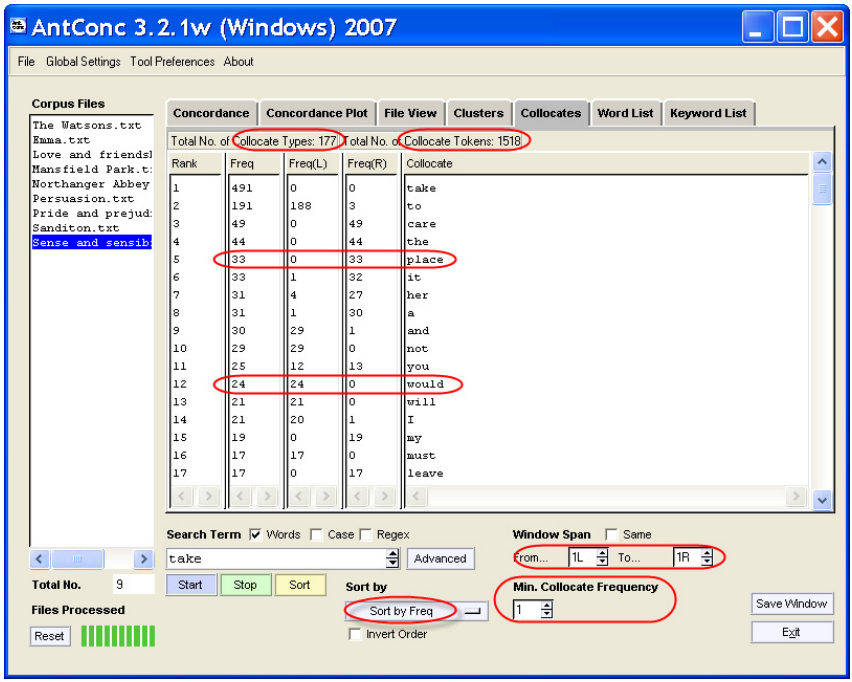


รูป 11.8 หน้าจอ File View

8.4 การหาคำปรากฏร่วม

ในหน้าจอ Collocate เป็นการแสดงคำปรากฏร่วมของคำหลัก Collocate Type คือจำนวนศัพท์ที่ปรากฏร่วม Collocate Token คือจำนวนคำปรากฏร่วมที่นับได้ทั้งหมด Window Span ใช้กำหนดขอบเขตหน้าหลังของคำหลักว่าต้องการดูคำปรากฏร่วมไกลเพียงใด 1L-1R คือดูคำติดกันข้างหน้าและข้างหลัง Min Collocate Frequency ใช้ตั้งเงื่อนไขว่าจะพิจารณาเฉพาะคำปรากฏร่วมที่มีความถี่อย่างน้อยเท่าใด ในตัวอย่างนี้ตั้งค่าเป็น 1 คือปรากฏครั้งเดียวก็นับ แต่ในกรณีใช้งานจริงกับคลังข้อมูลขนาดใหญ่ควรกำหนดไว้อย่างน้อย 5 การจัดเรียงในรูปเป็นการเรียงตามความถี่ของคำปรากฏร่วม ในตัวอย่างจะ

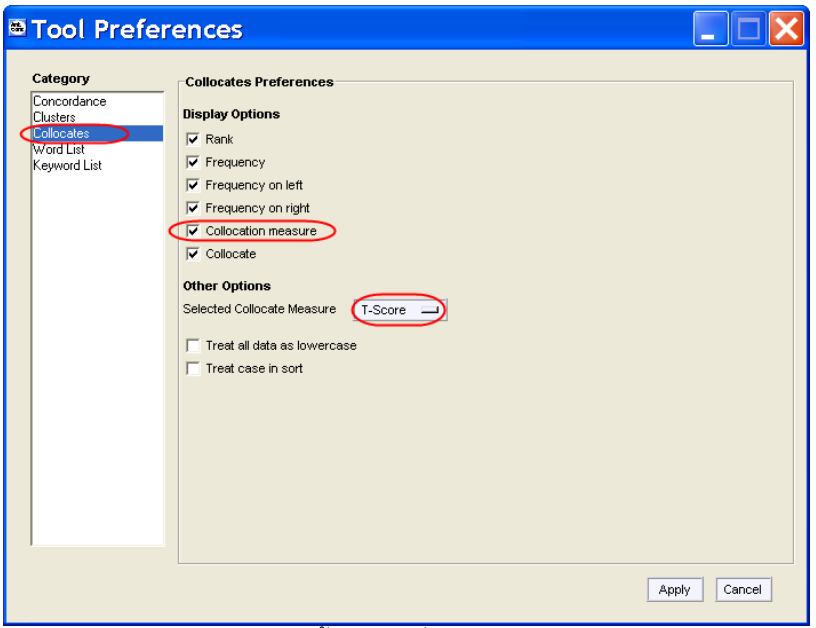
เห็น place ปรากฏร่วมหลัง take 33 ครั้งไม่มีปรากฏข้างหน้า ส่วนคำ would ปรากฏหน้า take 24 ครั้งไม่มีปรากฏข้างหลัง



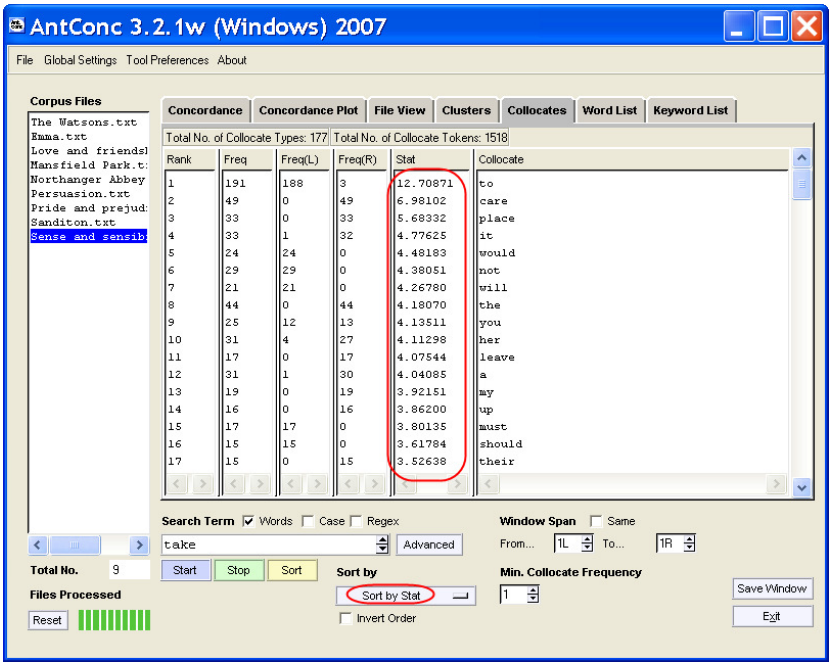
รูป 11.9 หน้าจอ collocate

การแสดงคำปรากฏร่วมโดยอาศัยความถี่อย่างเดียว มีผลให้คำที่ปรากฏบ่อยในข้อมูลขึ้นมาในลำดับต้นได้โดยไม่มีความสัมพันธ์ใดกับคำหลัก เช่น คำว่า the, a เป็นต้น เราสามารถสั่งให้แสดงคำปรากฏร่วมให้จัดลำดับใหม่โดยอาศัยการคำนวณทางสถิติ คำที่ปรากฏน้อยแต่มักปรากฏกับคำหลักเสมอๆ ก็ได้รับน้ำหนักมากขึ้นกว่าคำที่ปรากฏบ่อยแต่ปรากฏกับคำอื่นๆ ด้วยมาก วิธีใช้การคำนวณทางสถิตินี้ ต้องเลือกที่ Tool Preference ก่อน เลือก Collocates และเช็ค Collocate measure และเลือกสถิติที่ต้องการใช้ ซึ่งมีสองแบบ คือ MI

(mutual information) และ T-score หลังจากนั้นให้กลับไปหน้าจอ Collocate แล้วเลือก sort by stat แล้วกดปุ่ม Start ใหม่ โปรแกรมจะใช้เวลานานมากขึ้น เพราะต้องคำนวณผล เราสามารถปรับ Min Collocate Frequency เป็น 3 หรือ 5 ได้เพื่อให้ได้ผลที่เร็วขึ้น ผลที่ได้จะแสดงคำปรากฏร่วมที่ใกล้เคียงความจริงมากขึ้น คำว่า the, a จะตกลำดับไป



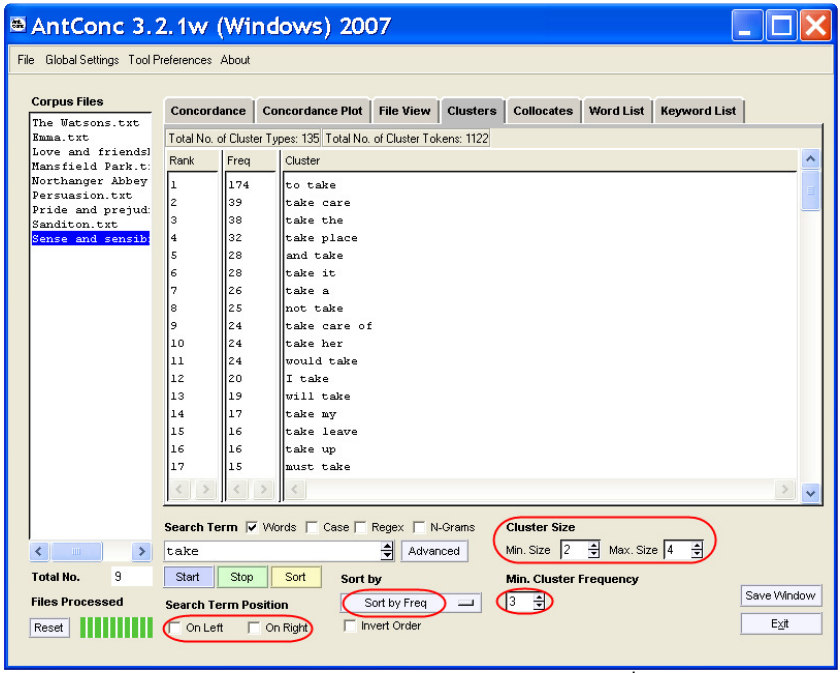
รูป 11.10 การตั้งค่าสถิติที่ใช้ในการหา collocate



รูป 11.11 หน้าจอ collocate เรียงตามค่าสถิติ

8.5 การค้นหากลุ่มคำ

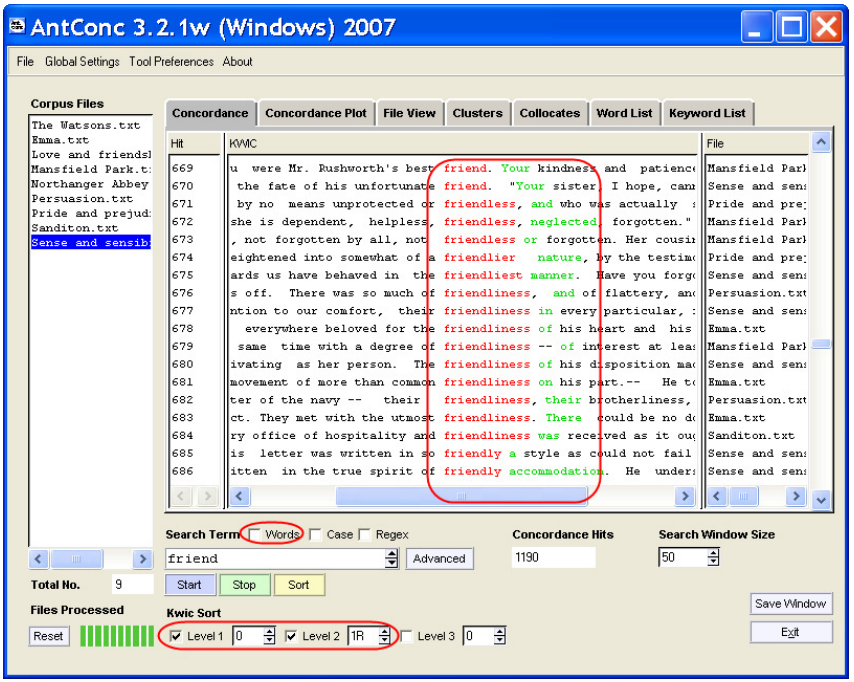
หน้าจอ Cluster ใช้ในการแสดงผลกลุ่มคำที่อยู่ติดกัน โดยมีคำหลักที่ต้องการค้นเป็นส่วนหนึ่ง เราสามารถกำหนดจำนวนกลุ่มคำว่าน้อยสุดและมากสุดเป็นเท่าไร ในที่นี้กำหนดหากลุ่มคำ 2 ถึง 4 คำ และกำหนดว่าแต่ละกลุ่มต้องพบอย่างน้อย 3 ครั้ง และให้จัดเรียงผลตามความถี่กลุ่มคำที่พบ หากต้องการหาเฉพาะกลุ่มคำที่มีคำหลักอยู่ซ้ายสุดหรือขวาสุดก็เลือกจาก Search Term Position



รูป 11.12 หน้าจอ clusters เรียงตามความถี่

8.6 การค้นแบบซับซ้อน

ในการค้นหาคำหลักนั้น นอกจากการค้นแบบคำเต็มที่ผ่านมา เรายังสามารถค้นแบบที่ซับซ้อนมากขึ้นได้อีก หากสังเกตช่อง Search term จะมีตัวเลือกต่างๆให้ใช้ การขีดถูกหน้าตัวเลือก Words หมายความว่าให้ค้นหาคำหลักนั้นโดยมองเป็นคำเต็มๆ แต่ถ้าเอาเครื่องหมายขีดถูกนี้ออกจะหมายความว่าคำหลักที่ค้นนั้นไม่จำเป็นต้องเป็นคำเต็ม สามารถเป็นส่วนหนึ่งของคำได้ ดังตัวอย่างการค้น friend ในข้างล่าง จะเห็นว่าเราจะได้คำอื่นอย่าง friendless, friendlier, friendliness, friendly เป็นต้น ซึ่งเมื่อคำหลักมีหลายรูป เวลาจัดเรียงเราควรจัดเรียงคำหลักเป็นอันดับแรกก่อน แล้วเลือกคำบริบทในลำดับต่อไป

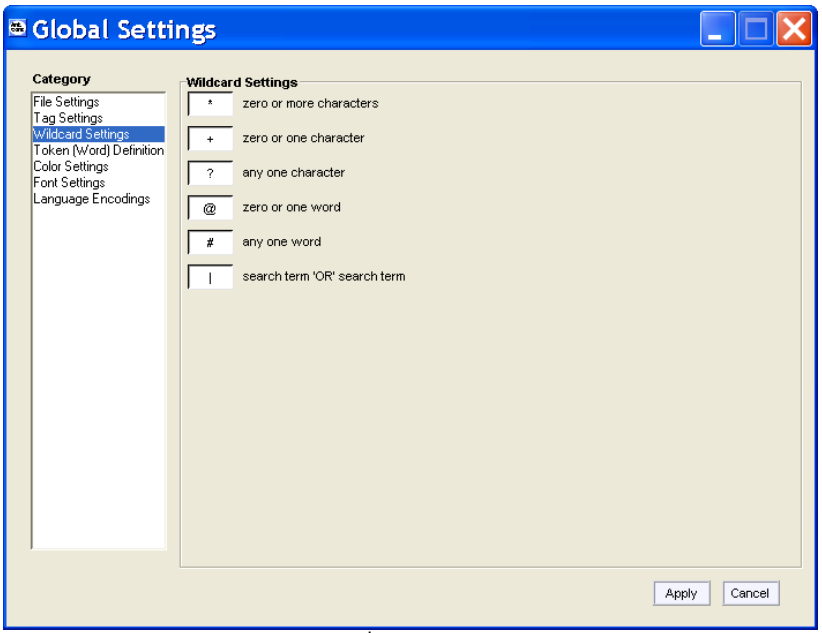


รูป 11.13 การค้น concordance โดยเทียบหาบางส่วนของคำ

โปรแกรม AntCon รองรับการค้นคำโดยใช้ wildcard character อย่าง * ? และการค้นโดยใช้ regular expression ความหมายของ wildcard character กำหนดไว้ใน Global Settings โดยที่ * แทนตัวอักษรอะไรกี่ตัวก็ได้หรือไม่มีเลยก็ได้ + แทนอักษรหนึ่งตัวหรือไม่มีเลย ? แทนตัวอักษรใดใดหนึ่งตัว @ แทนคำหนึ่งคำหรือไม่มีคำก็ได้ # แทนคำหนึ่งคำ | แทนความหมายของหรือ ตัวอย่างเช่น

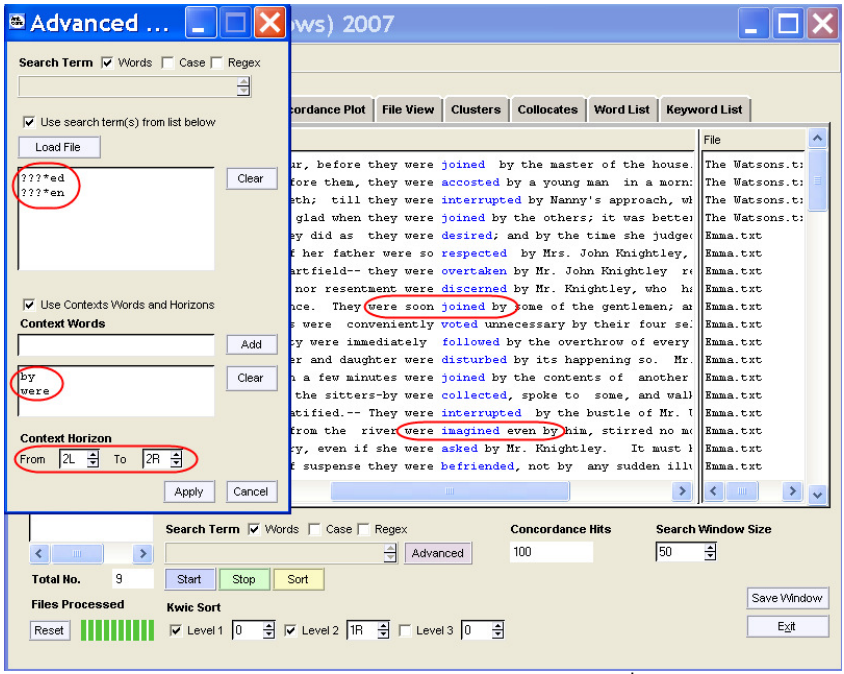
- เมื่อค้น un*ly หมายถึงคำใดใดที่ขึ้นต้นด้วย un และลงท้ายด้วย ly เช่น unlikely, undoubtedly, unfortunately, unhappily เป็นต้น

- เมื่อค้น very#thing จะได้คำสามคำที่มีคำใดใดระหว่าง very กับ thing เช่น very unlikely thing, very serious thing, very bad thing, very best thing, เป็นต้น
- เมื่อค้น break|broke|broken จะได้คำหลักทั้งสามคำนี้ในข้อมูล



รูป 11.14 สัญลักษณ์ที่ใช้ในการค้นหาแบบ wildcard

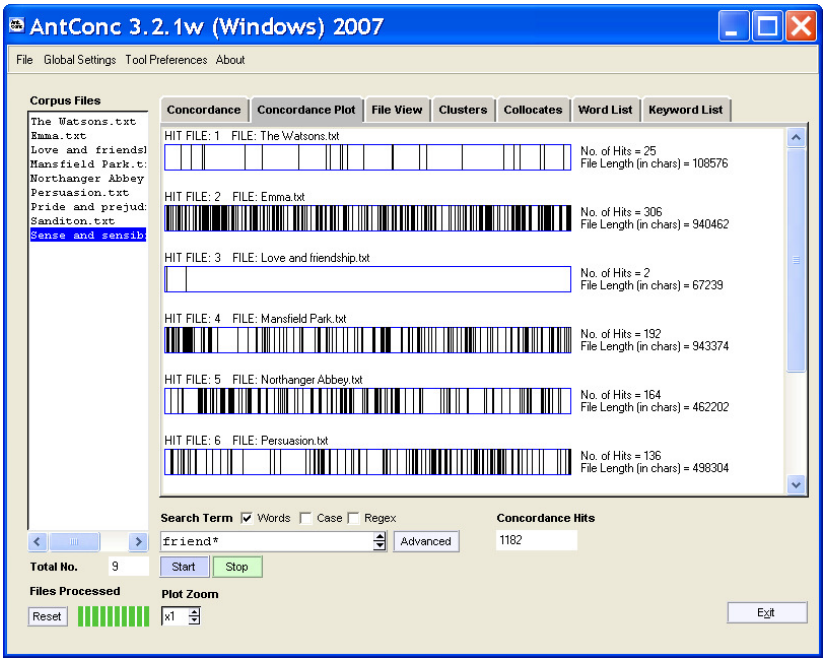
ในการค้นแบบ advanced mode เราสามารถค้นคำหลักหลายๆ คำ พร้อมกับกำหนดบริบทข้างเคียงว่าจะต้องมีคำใดบ้าง ในตัวอย่างข้างล่าง เป็นการค้นคำที่มีอย่างน้อยสามตัวอักษรแล้วลงท้ายด้วย ed หรือ en และในบริบทหน้าหลังสองตำแหน่ง 2L,2R จะต้องมีคำว่า were กับ by อยู่ ผลที่ได้ก็จะเป็นดังที่ปรากฏ



รูป 11.15 การค้นแบบซับซ้อนและกำหนดบริบทคำที่ต้องการ

8.7 การแสดงตำแหน่งการปรากฏของคำ

Concordance Plot เป็นการแสดงภาพการปรากฏของคำหลักที่ค้นว่าปรากฏ ณ ตำแหน่งใดในแต่ละไฟล์ที่ค้น เส้นขีดแต่ละเส้นแสดงตำแหน่งที่คำหลักปรากฏอยู่ Concordance Plot เป็นประโยชน์ในการมองภาพว่าคำหลักที่ค้นนั้นมีกล่าวถึง ณ ส่วนใดของเรื่อง ในตอนเริ่มต้น ตอนกลางเรื่อง หรือตอนท้ายเรื่อง เป็นต้น ในตัวอย่างข้างล่างเป็นการค้นคำ friend* คือทุกคำที่ขึ้นต้นด้วย friend ซึ่งจะเห็นว่านิยาย Emma จะมีการใช้คำพวกนี้มาก ในขณะที่นิยาย love and friendship กลับมีการใช้คำในกลุ่ม friend นี้น้อยมาก

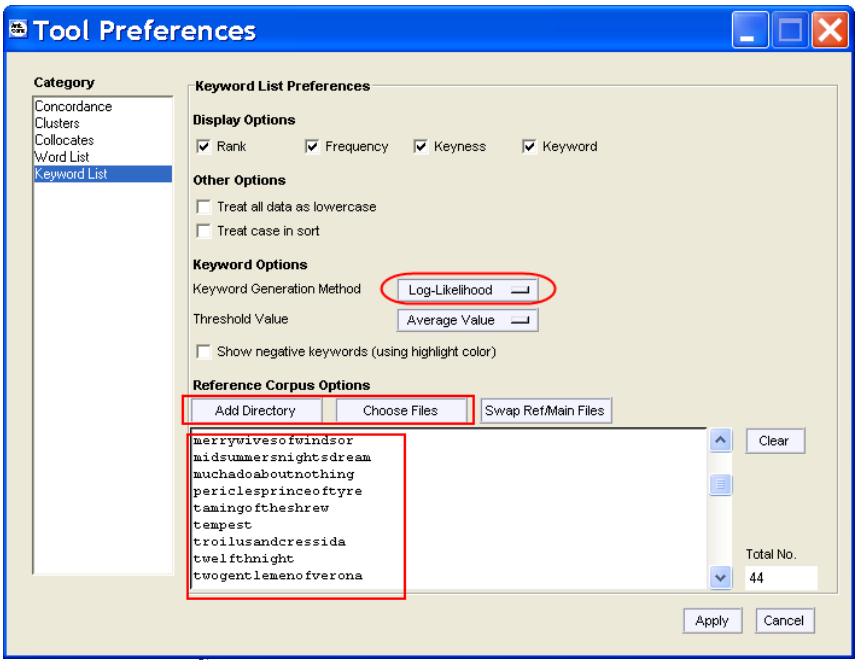


รูป 11.16 หน้าจอแสดง concordance plot

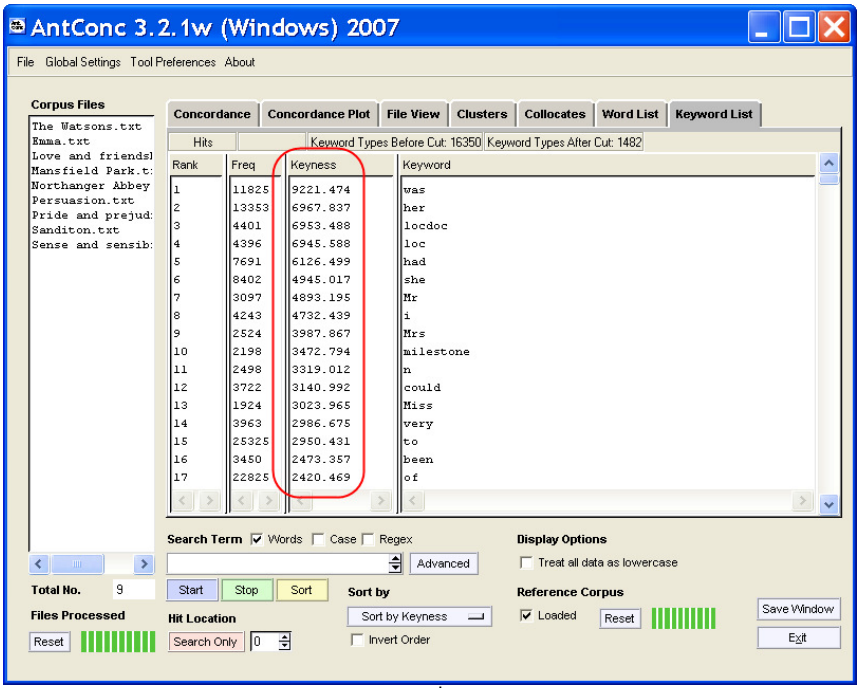
8.8 การหาคำสำคัญเมื่อเทียบกับคลังข้อมูลอ้างอิง

การใช้ Keyword List เป็นฟังก์ชั่นพิเศษที่ AntConc มีให้ไว้เพื่อเปรียบเทียบระหว่างคลังข้อมูลสองคลังข้อมูล เพื่อหาคำที่มีความสำคัญในคลังข้อมูลหนึ่งมากกว่าเมื่อนำไปเปรียบเทียบกับอีกคลังข้อมูลหนึ่งที่เรียกว่าคลังข้อมูลอ้างอิง หลักการโดยทั่วไปเป็นการมองหาคำซึ่งพบมากในคลังข้อมูลที่วิเคราะห์อยู่แต่พบน้อยในคลังข้อมูลอ้างอิง การที่จะใช้ส่วนนี้ได้ เราจึงต้องมีคลังข้อมูลอ้างอิงหรือเรียกว่า reference corpus ซึ่งสามารถกำหนดได้ใน Tool

Preferences ของ Keyword List โดยเลือก Add Directory หรือ Files ก็ได้ การเปรียบเทียบความเป็นคำสำคัญอาศัยวิธีการคำนวณทางสถิติซึ่งในโปรแกรมนี้มีไว้เลือกใช้ได้สองแบบ คือ Log-Likelihood หรือ Chi-square เมื่อจัดเรียงผลตาม keyness เราจะได้คำที่เป็นคำเฉพาะของคลังข้อมูลที่ศึกษาอยู่เรียงตามค่า keyness ทางสถิติ



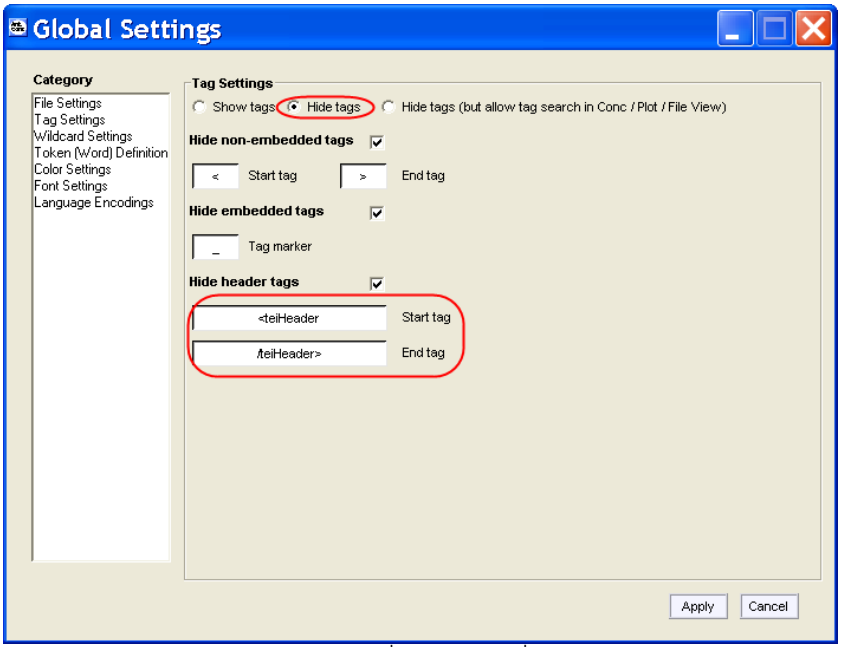
รูป 11.17 การตั้งค่าสถิติและ reference corpus สำหรับการหา keyword



รูป 11.18 หน้าจอแสดง keyword เมื่อเทียบกับ reference corpus

8.9 การใช้ AntConc กับข้อมูล XML

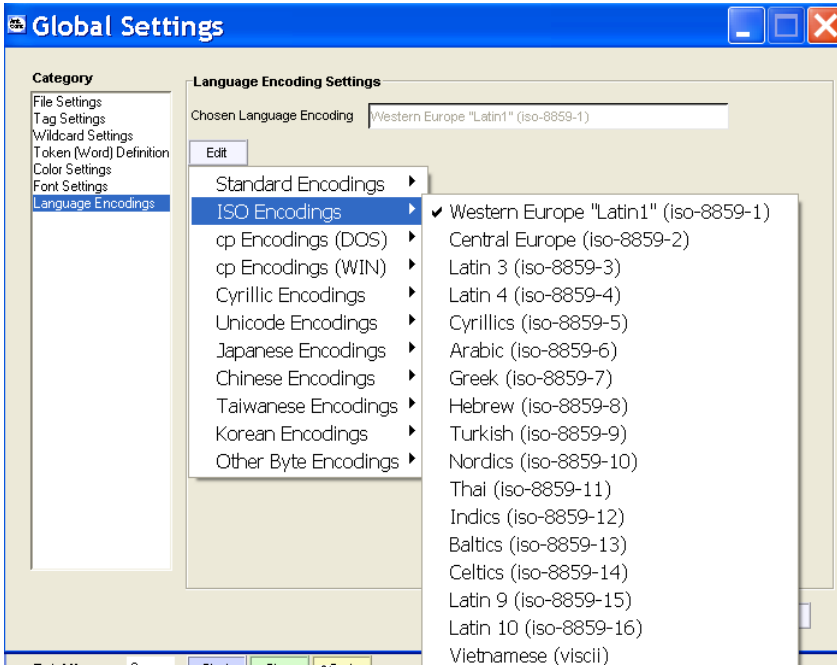
การใช้งานคลังข้อมูลที่มีการกำกับข้อมูลแบบ XML, SGML ตามมาตรฐานของ TEI เช่นข้อมูลที่โหลดจาก Oxford text archive, British National Corpus ฯลฯ โดยเราสามารถ ignore tag ต่างๆที่กำกับไว้ในคลังข้อมูลได้ รวมถึง ignore header tag ทั้งหมด ค่าตั้งต้นของ header tag คือ teiHeader แต่หากคลังข้อมูลที่ใช้ใช้ header tag ในชื่ออื่นก็ต้องปรับแก้ให้ตรงกับข้อมูลที่ใช้



รูป 11.19 การกำหนดเกี่ยววกับแท็กที่กำกับในข้อมูล

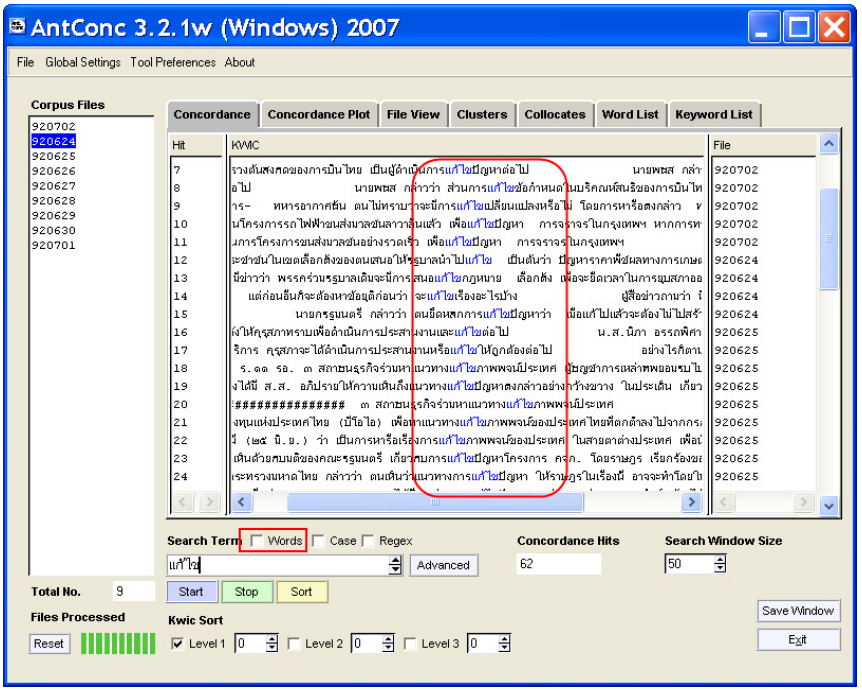
8.10 การใช้ AntConc กับข้อมูลภาษาไทย

โปรแกรม AntConc นี้เป็นโปรแกรมที่พัฒนาโดย Laurence Anthony ที่มหาวิทยาลัย Waseda ประเทศญี่ปุ่น จึงสามารถรองรับภาษาอื่นๆ ที่ไม่ใช่ภาษาอังกฤษได้หลากหลาย เช่น ภาษาญี่ปุ่น ภาษาจีน ภาษาเกาหลี ภาษาไต้หวัน ตลอดจนภาษาต่างๆ ที่เข้ารหัสอักขระแบบ ISO ผู้ใช้ต้องเลือกรหัสอักขระที่ตรงกับข้อมูลของตนที่จะใช้ ในกรณีภาษาไทย เราสามารถเลือก ISO Encoding ของภาษาไทยคือ ISO-8859-11 ได้



รูป 11.20 การตั้งค่าสำหรับใช้กับข้อมูลภาษาไทย

อย่างไรก็ตาม เนื่องจากข้อความภาษาไทยไม่มีการแบ่งคำ จึงไม่สามารถค้นได้เต็มประสิทธิภาพเหมือนการค้นภาษาอังกฤษ การค้นคำไทยจะต้องเลือกให้เป็นส่วนหนึ่งของคำโดยค้นแบบไม่เต็มคำ และผลที่แสดงหน้าจอจะไม่เป็นระเบียบเรียบร้อยตรงแนวเหมือนอย่างภาษาอังกฤษ และไม่สามารถจัดเรียงลำดับตามบริบทซ้ายหรือขวาได้ โปรแกรม AntConc จึงใช้กับข้อมูลภาษาไทยได้เพียงในระดับหนึ่ง



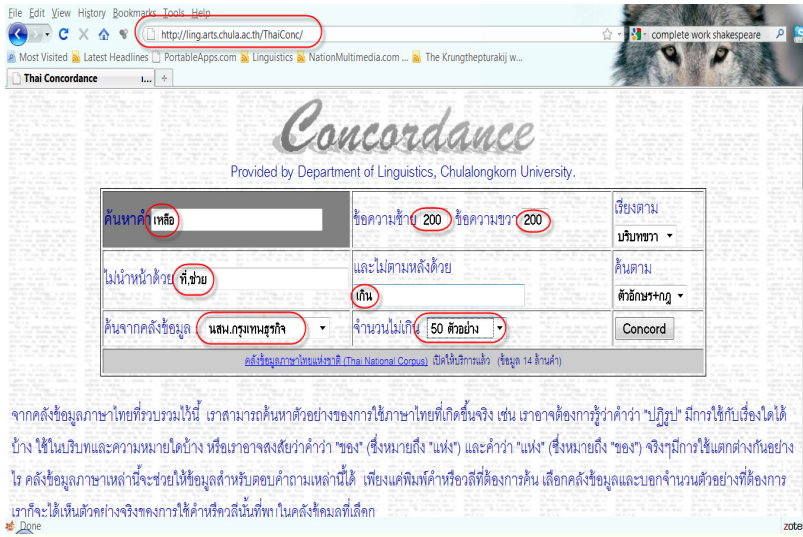
รูป 11.21 ผลแสดงการค้น concordance ข้อมูลภาษาไทย

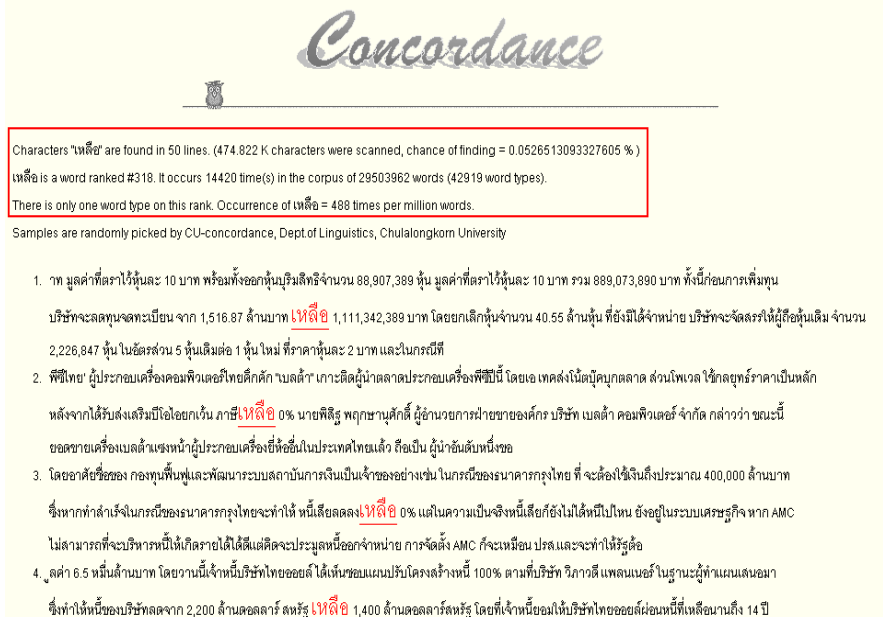
9. การใช้โปรแกรมคอนคอร์แด็นซ์ภาษาไทย

โปรแกรม Thai Concordance Online เป็นบริการค้นหาข้อมูลตัวอย่างการใช้ภาษาไทยผ่านทางอินเทอร์เน็ตที่ http://ling.arts.chula.ac.th/ThaiConc/ ซึ่งเป็นบริการที่ให้เลือกค้นคำจากคลังข้อมูลภาษาไทยที่กำหนดให้จำนวนหนึ่ง เช่น หนังสือพิมพ์ รายงานข่าว บทความวิชาการ เป็นต้น ผู้ใช้สามารถกำหนดความยาวของบริบทซ้ายขวาที่ต้องการว่าเป็นกี่ตัวอักษร พร้อมทั้งระบุคำข้างหน้าหรือข้างหลังคำหลักว่าไม่ต้องการให้เป็นคำอะไรบ้าง ทั้งนี้ก็เพื่อช่วยลดข้อมูลที่ไม่

ต้องการออกไป เช่น ถ้าเราต้องการค้นคำว่า เหลือ ที่เป็นกริยา เราก็จะไม่ต้องการคำนาม “ที่เหลือ” คำกริยา “ช่วยเหลือ” หรือคำวิเศษณ์ “เหลือเกิน” เป็นต้น ผลที่ได้จากโปรแกรมเป็นการดึงตัวอย่างตามจำนวนที่ระบุให้ดึงออกมา ซึ่งการค้นแต่ละครั้งจะได้ตัวอย่างที่ต่างกัน เพราะเป็นการค้นแบบสุ่ม

ที่บรรทัดบนสุดเป็นรายงานผลการค้นคำหลักที่ระบุ โดยบอกว่าเพื่อให้ได้ตัวอย่าง 50 ตัวอย่างที่ต้องการ โปรแกรมได้อ่านข้อความต่อเนื่องไปกี่ตัวอักษรและเมื่อคำนวณเป็นความน่าจะเป็นที่จะพบคำว่า เหลือ ในคลังข้อมูลจะเป็นเท่าไร นอกจากนี้ ยังให้ข้อมูลเพิ่มเติมว่า คำว่า”เหลือ” นี้เป็นคำที่พบมากเป็นลำดับที่ 318 ซึ่งมีจำนวน 14,420 ครั้งในคลังข้อมูลที่เลือกซึ่งมีขนาด 29,503,962 คำ และคลังข้อมูลนี้มีศัพท์42,919 คำ คำในตำแหน่งที่ 318 นี้มีเพียงคำเดียว และหากคำนวณการปรากฏของคำ เหลือ จะได้ว่าในทุกหนึ่งล้านคำจะพบคำนี้ 488 ครั้ง ข้อมูลเหล่านี้จะช่วยให้เห็นว่า คำหลักที่ค้นนั้นปรากฏมากน้อยเพียงใด





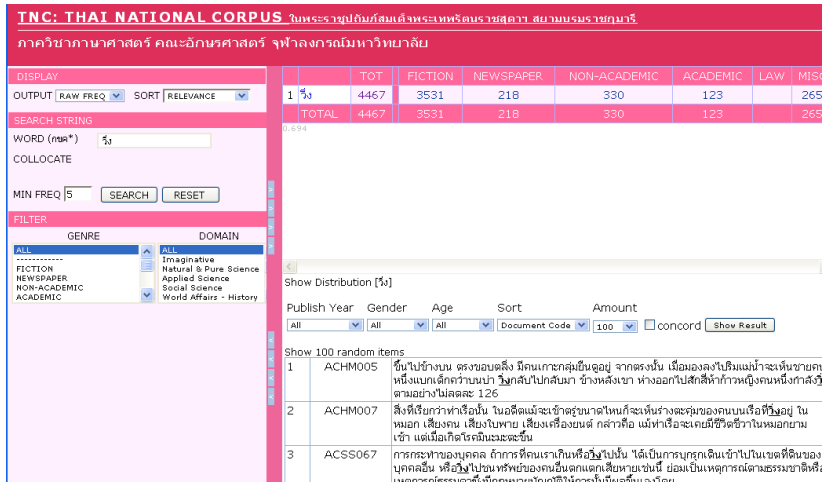
รูป 11.22 หน้าเว็บ Thai Concordance และผลจากการค้น

10. การใช้คลังข้อมูลภาษาไทยแห่งชาติ

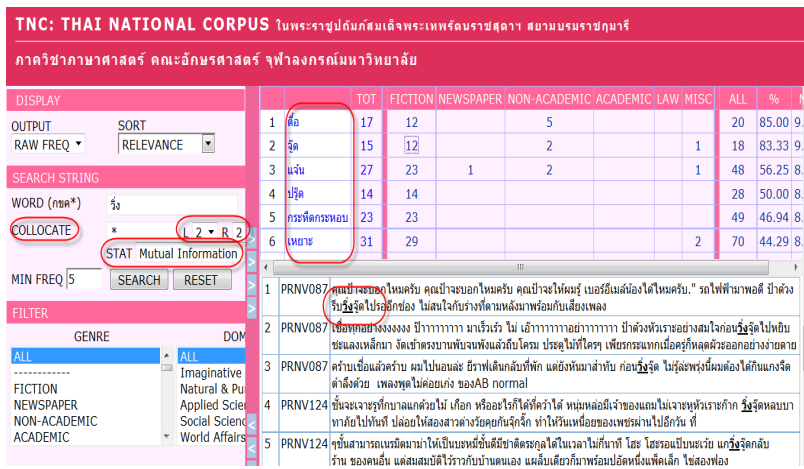
คลังข้อมูลภาษาไทยแห่งชาติเป็นคลังข้อมูลภาษาแบบทั่วไป มีการเก็บข้อมูลภาษาเขียนที่หลากหลายจากแหล่งต่างๆ ให้ครอบคลุมการใช้ภาษาไทยในลักษณะต่างๆ ถูกออกแบบมาโดยเทียบเคียงกับโครงสร้างคลังข้อมูลภาษาเขียนของ British National Corpus ปัจจุบันมีข้อมูล 18 ล้านคำ สามารถสืบค้นคลังข้อมูลได้ที่ http://ling.arts.chula.ac.th/TNC/

ผู้ใช้สามารถระบุคำที่ต้องการค้น และระบุประเภทตัวบท (genre) และหรือเนื้อหา (domain) ผลที่ได้จะแจกแจงความถี่โดยรวมและความถี่ที่แยกไปตามตัวบทประเภทต่างๆ ได้แก่ นิยาย หนังสือพิมพ์ กึ่งวิชาการ วิชาการ กฎหมาย และอื่นๆ ผู้ใช้สามารถเลือกกดที่ตัวเลขความถี่เหล่านี้ ก็จะได้รายการตัวอย่างการใช้คำนั้นตามจำนวนที่ระบุ หลังจากนั้นสามารถกรองผลที่ได้โดยกำหนดตัวกรองเป็นปี เพศ ช่วงอายุผู้เขียน และสามารถเลือกแสดงผลแบบจัดคำหลักไว้กลางได้ รหัสข้างหน้าตัวอย่างบอกถึงแหล่งที่มาของข้อมูล ซึ่งผู้ใช้สามารถกดดรายละเอียดได้ พร้อมทั้งจะเห็นบริบทที่มากขึ้นของตัวอย่างนั้น

นอกจากผลการค้นแบบคำหลักพร้อมบริบท ในเว็บ TNC ยังสามารถค้นข้อมูลคำปรากฏร่วมได้ด้วย ซึ่งใน Thai Concordance Online ไม่สามารถให้ข้อมูลนี้ได้ โดยเมื่อกดที่ collocate ก็จะมีตัวเลือกสถิติที่ต้องการใช้และการระยะห่างของคำ เมื่อได้รายการคำที่เป็น collocate แล้ว เราสามารถกดที่คำที่ต้องการก็จะได้ตัวอย่างการใช้คำหลักนั้นพร้อมคำปรากฏร่วม



รูป 11.23 หน้าเว็บ TNC และผลจากการสืบค้น concordance



รูป 11.24 หน้าเว็บ TNC และผลจากการสืบค้น collocate

บทที่ 12 คำปรากฏร่วมและหน่วยเสมือนวลี

1. ความเป็นมา

ในภาษาใดก็ตาม เรามักจะสังเกตพบว่าคำบางคำมักจะปรากฏใช้ร่วมกับคำบางคำอยู่เสมอ เช่น ในภาษาอังกฤษที่ *pay* มักปรากฏใช้ร่วมกับ *attention*, หรือคำว่า *look* มักปรากฏร่วมกับ *after* เป็นปกติอย่างเห็นได้ชัด การที่คำบางคำมักปรากฏใช้ร่วมกันอยู่เสมอนี้ เราเรียกว่าคำปรากฏร่วม collocation Firth (1957:6) เป็นคนแรกๆ ที่ใช้คำว่า collocation และให้นิยามของการปรากฏร่วมกันของคำว่า หมายถึง คำหรือวลีที่มักปรากฏใช้ร่วมกันจนเป็นปกติ (collocation of a given word are statements of the habitual or customary places of that word) Fontenelle (1994) ได้ยกตัวอย่างของคำว่า *sour* กับ *rotten* คือ เราจะพบคำว่า *sour* จะปรากฏร่วมกับ *milk (sour milk)* มากกว่าที่จะพบคำว่า *rotten* ปรากฏร่วมกับ *milk (rotten milk)* ถึงแม้ว่าการใช้ *rotten milk* จะไม่ผิดหลักไวยากรณ์ แต่เป็นการใช้ที่ค่อนข้างแปลกสำหรับผู้พูดภาษาอังกฤษ คำว่า *rotten* นั้นจะใช้ร่วมกับ *egg* มากกว่า ตัวอย่างนี้ แสดงให้เห็นว่าในการใช้ภาษาจริง ผู้พูดเลือกที่จะใช้คำบางคำร่วมกับคำบางคำเป็นพิเศษ ซึ่งสิ่งนี้ไม่สามารถอธิบายโดยกฎทางวากยสัมพันธ์ได้ (เพราะทั้งคำ *sour* และ *rotten* ต่างก็เป็นคำคุณศัพท์) ระดับของการปรากฏร่วมกันของคำนั้น มีตั้งแต่คำที่ปรากฏอยู่ด้วยกันเสมอ จนมีลักษณะเป็นเสมือนสำนวน (idiom) เช่น *kick the bucket* ซึ่งมีความหมายว่าตาย หรือ *to lick somebody's boot* ซึ่งมีความหมายว่าประจบสอพลอ ไปจนถึงระดับที่คำสามารถปรากฏร่วมกันอย่างอิสระ คือ เป็นไปตามกฎไวยากรณ์ว่าสามารถใช้ร่วมกันได้ เช่น *kick a football* หรือ *fire staff* เพราะ

สามารถแทนที่คำ *a football* หรือ *fire* ได้หลากหลาย เช่น *kick a cat* หรือ *lay off staff* ในกรณีของคำที่เป็นเสมือนสำนวนนั้น นักภาษาศาสตร์ส่วนใหญ่ไม่จัดว่าเป็นคำปรากฏร่วม เพราะสื่อถึงหน่วยความหมายเพียงหน่วยเดียว (a single semantic entity) ความหมายของวลีนั้นไม่สามารถวิเคราะห์จากความหมายของคำย่อยๆ ภายใน และในกรณีของการปรากฏร่วมกันได้อย่างอิสระก็ไม่ถือเป็นเรื่องของคำปรากฏร่วมเช่นกัน คำปรากฏร่วมจึงเป็นกรณีที่อยู่ระหว่างทั้งสองขั้วนี้ แต่ถึงกระนั้น ขอบเขตที่กำหนดนี้ก็ยังกว้าง ความหมายของคำปรากฏร่วมจึงไม่เป็นที่ตกลงกันแน่นอนว่าครอบคลุมขอบเขตมากน้อยเท่าใด

นอกจากคำ คำปรากฏร่วม (collocation) คำอีกจำพวกหนึ่งซึ่งถูกกล่าวถึงและเป็นประเด็นที่มีผู้สนใจศึกษาอย่างมาก คือคำว่า หน่วยเสมือนวลี (phraseological unit) หรือหน่วยประกอบหลายคำ (multi-word unit) ซึ่งเป็นหน่วยที่ดูเสมือนเป็นวลีแต่โดยลักษณะการใช้แล้วถือเป็นหนึ่งหน่วยภาษาที่ไม่ควรมองแยกองค์ประกอบภายใน ทั้งคำปรากฏร่วมและหน่วยเสมือนวลีนี้มีความคล้ายคลึงกันและเป็นที่สนใจศึกษามากในปัจจุบัน บางคนก็ใช้คำหน่วยเสมือนวลีในความหมายกว้างที่คลุมคำปรากฏร่วมด้วย ในปีค.ศ.2005 มีการประชุมวิชาการของผู้ที่สนใจเรื่องนี้มาจากศาสตร์ต่างๆกว่า 170 คน คือการประชุม Phraseology 2005: The Many Faces of Phraseology ประกอบด้วยนักวิชาการด้านภาษาศาสตร์คลังข้อมูล นักพจนานุกรม นักวิจัยการประมวลผลภาษา นักภาษาศาสตร์จิตวิทยา นักวิชาการด้านการสอนภาษา นักทฤษฎีภาษาศาสตร์ ฯลฯ ทั้งหมดมีความสนใจในปรากฏการณ์เดียวกันของภาษาแต่จากแง่มุมที่แตกต่างกัน ความสนใจเรื่องนี้มีมากจนสามารถแยกเป็นแขนงการศึกษาหนึ่งที่เรียกว่า phraseology ผลจากการประชุมวิชาการในครั้งนั้นได้มีหนังสือออกมาสามเล่ม เล่มแรกเป็นภาษาฝรั่งเศส La phraséologie dans tous

ses états เล่มที่สอง Phraseology: An Interdisciplinary Perspective (2008) รายงานความสนใจในหน่วยเสมือนวลีในศาสตร์ต่างๆ และเล่มที่สาม Phraseology in Foreign Language learning and Teaching (2008) ให้ความสนใจในเรื่องของหน่วยเสมือนวลีในการเรียนการสอนภาษาเป็นหลัก

Ellis (2008) กล่าวสรุปความเป็นมาในเรื่องนี้ว่า ความสนใจในหน่วยเสมือนวลีมีมาตั้งแต่ในยุคไวยากรณ์โครงสร้าง นักไวยากรณ์โครงสร้างอย่าง Fries (1952) ได้แยกความแตกต่างระหว่างความหมายจากคำ (lexical meaning) และความหมายจากโครงสร้าง (structural meaning) ซึ่งได้มาจากรูปแบบโครงสร้างบางประเภท การเรียนภาษาจึงไม่ใช่เพียงแค่รู้ความหมายของคำ แต่ต้องเรียนรู้รูปแบบการประกอบคำและความหมายของโครงสร้างนั้น Harris (1982, 1991) นักภาษาศาสตร์โครงสร้างได้ใช้แบบจำลองคณิตศาสตร์เสนอไวยากรณ์ตัวดำเนินการ (Operator Grammar) ซึ่งอธิบายว่าภาษาเป็นระบบที่คุณสมบัติด้านวากยสัมพันธ์และอรรถศาสตร์ของคำต่างๆ จะต้องอธิบายในเชิงสัมพันธ์กับคำอื่นๆ และการเรียนรู้ภาษาเป็นเรื่องของการเรียนรู้ผ่านการใช้ในบริบทสังคม ความคิดของนักไวยากรณ์โครงสร้างเหล่านี้ ได้ถูกลดทอนความสนใจในยุคไวยากรณ์เพิ่มพูนที่มองไวยากรณ์แบบแยกขาดจากคลังศัพท์อย่างชัดเจน คำเสมือนวลีพวกนี้ถูกจัดเป็นข้อมูลชายขอบไม่ใช่เรื่องหลักของภาษาศาสตร์ แต่ในปัจจุบัน ความสนใจเรื่องรูปแบบและความหมายของโครงสร้างของนักไวยากรณ์โครงสร้างเหล่านี้มาปรากฏในไวยากรณ์ปริชานไวยากรณ์หน่วยสร้าง ที่มองหน่วยสัญญะ (symbolic unit) เป็นสำคัญซึ่งเป็นได้ทั้งคำและหน่วยที่ใหญ่กว่าคำ

การเติบโตของภาษาศาสตร์คลังข้อมูลมีส่วนสำคัญที่ทำให้ความสนใจหน่วยที่ใหญ่กว่าคำเป็นที่แพร่หลาย เพราะด้วยการใช้ข้อมูลภาษาจำนวนมาก

และวิธีการทางสิถิติต่างๆ ทำให้ภาษาศาสตร์คลังข้อมูลสามารถสร้างเครื่องมือที่จะช่วยดึงหน่วยเหล่านี้ออกมาให้ศึกษาได้โดยง่าย

2. ประเภทของคำปรากฏร่วม

Nesselhauf (2005) มองว่าคำปรากฏร่วมเป็นปรากฏการณ์ทางภาษาที่อยู่ระหว่างสองขั้ว คือการปรากฏแบบอิสระ (free combination) กับแบบสำนวนตายตัว (idiom) สำนวนตายตัวอย่าง *sweeten the pill* มีความหมายเฉพาะว่าพยายามทำให้รู้สึกดีขึ้นในภาวะที่ได้ข่าวร้ายมา เราจะไม่สามารถแยกเอาเฉพาะคำอย่าง *pill* ไปใช้กับกริยาอื่นในความหมายข่าวร้ายนี้ได้ *(*I was confronted with a pill*) หรือ *kick the bucket* ก็จะมีความหมายเฉพาะว่า ตาย ไม่สามารถแยกส่วนไปใช้ในความหมายนี้ได้ ส่วนการปรากฏอิสระอย่าง *want a car* สามารถเปลี่ยนคำจาก *car* เป็นนามคำอื่นๆ ได้ หรือเปลี่ยนกริยา *want* เป็น *wash* ก็ได้ ส่วนคำปรากฏร่วมอย่าง *take a picture* แม้ว่าเราอาจเปลี่ยนคำ *picture* เป็น *photograph* ได้แต่ก็ไม่สามารถเปลี่ยนเป็นคำอื่นๆ ได้ทั้งหมด เราไม่ใช้ *take a film* หรือ *movie*

คำปรากฏร่วมโดยทั่วไปจะประกอบด้วยคำเนื้อหา (open class) สองคำขึ้นไป เช่น *strong tea, powerful car* กรณีนี้ Cowie (1992) เรียกว่า lexical collocation เพราะแสดงถึงความสัมพันธ์เฉพาะที่มีอยู่ระหว่างคำเหล่านั้นโดยทั่วไปแล้ว ความหมายของกลุ่มคำที่เป็นคำปรากฏร่วมจะไม่เท่ากับความหมายที่เกิดจากการรวมความหมายของคำแต่ละคำ เช่น *strong tea* *strong* ในที่นี้ไม่ได้หมายถึงการมีสภาพทางกายที่แข็งแรงแต่หมายถึงรสชาติที่แก่ lexical collocation จะแตกต่างจาก grammatical collocation โดยที่ในกรณีหลังนี้จะมีคำหนึ่งเป็นคำไวยากรณ์ (closed class) เช่น *depend on, account for,*

rely on เป็นต้น ในกรณีของคำปรากฏร่วมที่ค่อนไปทางที่เป็นรูปตายตัว แต่ไม่ถึงขนาดเป็นเสมือนสำนวน Cruse (1986) เรียกคำปรากฏร่วมประเภทนี้ว่า bound collocation เช่น *pay attention* เพราะเราไม่สามารถแทรกคำชี้เฉพาะระหว่างคำทั้งสองนี้ เช่น **He paid his attention to ...* แต่ก็สามารถแทรกคำคุณศัพท์ได้ เช่น *He paid little attention to ...* อย่างไรก็ตามการแยกลักษณะนี้ ก็ไม่มีเกณฑ์ที่แน่ชัดว่าจะตัดสินว่ากลุ่มคำใดเป็น bound collocation ได้อย่างไร เพราะเป็นเรื่องของระดับว่าคำในกลุ่มนั้นมีความสัมพันธ์กันมากน้อยเพียงใด ในที่นี้เราจึงเลือกใช้คำกว้าง ๆ ว่า collocation หรือคำปรากฏร่วมสำหรับการที่คำบางคำมักปรากฏใช้ร่วมกับคำบางคำเป็นพิเศษ

3. เกณฑ์การพิจารณาคำปรากฏร่วม

Manning and Schutze (1999: 171) ได้สรุปถึงเกณฑ์หลักในการพิจารณาคำปรากฏร่วมที่มีการใช้อยู่ในงานต่าง ๆ ว่ามีสามประการ ดังนี้

1. Non-compositionality คือ ความหมายของคำปรากฏร่วมจะไม่ได้มาจากการรวมความหมายของส่วนย่อยๆ โดยตรง ในกรณีสุดขั้วของคำที่ใช้เป็นสำนวน (idiom) ความหมายของคำที่ได้อาจแตกต่างจากความหมายเดิมไปเลย เช่น *kick the bucket* ในกรณีของคำปรากฏร่วมนั้น จะมีการเสริมความหมายบางอย่างเข้ามาซึ่งไม่สามารถคาดเดาได้จากคำใดในกลุ่มนั้น เช่น *white wine, white woman* คำว่า *white* ในที่นี้ไม่ได้มีความหมายถึงสีขาวเหมือนกับคำว่า *white* ปกติ ลักษณะที่เป็น non-composite นี้บางคนก็จะเรียกว่าเป็น formulaic เป็นลักษณะปกติที่พบในภาษาว่าเราจะจำรูปแบบคำเป็นกลุ่มไว้ใช้เลยโดยไม่ต้องผ่านกระบวนการแยกองค์ประกอบ (Sinclair เรียกหลักการนี้ว่า idiom principle)

2. Non-substitutability คือ เราไม่สามารถแทนคำในคำปรากฏร่วมด้วยคำใกล้เคียงที่น่าจะใช้แทนกันได้ (near-synonym) เช่น ไม่สามารถใช้ *yellow wine* แทน *white wine* ถึงแม้ว่าสีเหลืองจะดูใกล้เคียงกับสีที่แท้จริงของไวน์ชนิดนี้ก็ตาม

3. Non-modifiability คือ เราไม่สามารถแก้ไขคำปรากฏร่วมโดยการเติมหน่วยคำ (lexical item) หรือผ่านกระบวนการทางไวยากรณ์ (grammatical transformation) ได้ โดยเฉพาะคำที่เป็นวลีตายตัวแล้ว (frozen expression) เช่น ไม่สามารถแก้ไขสำนวนอย่าง *to get a frog in one's throat* โดยการเติม *ugly* ให้เป็น *to get an ugly frog in one's throat* (เกณฑ์ข้อนี้สะท้อนลักษณะของกลุ่มคำที่เป็นเสมือนสำนวนมากกว่า เกณฑ์นี้จึงอาจจะเหมาะสำหรับคำปรากฏร่วมที่ค่อนไปในทางเป็นสำนวนมากกว่าคำปรากฏร่วมแบบปกติ)

นอกจากเกณฑ์ทั้งสามแล้ว ความถี่การปรากฏก็เป็นอีกปัจจัยหนึ่งในการพิจารณา เพราะแนมโน้มที่คำกล่มหนึ่งจะปรากฏใช้ร่วมกัน หากค้นในคลังข้อมูลภาษาก็น่าจะพบจำนวนการใช้มากกว่าปกติที่ควรเป็น

4. การระบุหาคำปรากฏร่วม

Manning and Schutze (1999: 173) เองก็ได้เสนอวิธีง่ายๆ แบบหนึ่งที่อาจใช้ทดสอบว่าคำนั้นๆเป็นคำปรากฏร่วมหรือไม่ โดยให้ทดลองแปลเป็นอีกภาษาหนึ่ง ถ้าไม่สามารถแปลทุกคำในลักษณะที่เป็นการแปลแบบคำต่อคำได้ ก็ให้สงสัยไว้ก่อนว่า คำชุดนั้นอาจเป็นคำปรากฏร่วมได้ เช่น คำว่า *make a decision* ซึ่งเมื่อแปลเป็นภาษาฝรั่งเศสแบบคำต่อคำจะได้ `faire une décision` ซึ่งไม่ใช่คำแปลที่ถูกเพราะว่าควรจะแปลเป็น `prendre une décision` มากกว่า ดังนั้น *make a decision* จึงน่าสงสัยว่าเป็นคำปรากฏร่วม

นอกจากการพิจารณาคำปรากฏร่วมตามมุมมองของนักภาษาศาสตร์แล้ว ก็มีผู้เสนอให้มองคำปรากฏร่วมจากมุมมองทางสถิติ โดยใช้วิธีการศึกษาจากคลังข้อมูลภาษา แล้วหาว่าคำที่มักปรากฏใช้ร่วมกันมากจนเกินกว่าที่จะเป็นเหตุบังเอิญนั้นถือเป็นคำปรากฏร่วมในทางสถิติ (probabilistic collocation) ซึ่งอาจจะต่างไปจากคำปรากฏร่วมในทางภาษาศาสตร์ (linguistic collocation) ก็ได้ บางคนจึงเรียกคำปรากฏร่วมในทางสถิติว่าเป็น co-occurence

การหาคำปรากฏร่วมได้จะเป็นประโยชน์ต่อการให้คำนิยามของคำนั้นในพจนานุกรม และในการจัดทำพจนานุกรมคำปรากฏร่วม (collocation dictionary) นอกจากนี้ Altenberg and Eeg-Olofsson (1990) ยังกล่าวว่าความรู้เรื่องคำปรากฏร่วมนั้นจำเป็นสำหรับการเรียนภาษาที่สอง ทั้งนี้เพราะข้อผิดพลาดของนักเรียนจำนวนมากสามารถอธิบายได้จากเรื่องของการใช้คำปรากฏร่วมว่ามีสาเหตุจากการที่นักเรียนไม่รู้ว่าจะต้องใช้คำอะไรร่วมกับคำที่ต้องการจึงจะถูกต้องเหมาะสม การนำแนวคิดด้านคำปรากฏร่วมในทางสถิติมาใช้ แม้ผลที่ได้อาจจะไม่ใช่คำปรากฏร่วมในทางภาษาศาสตร์เสมอไป แต่ก็เป็นวิธีการที่สะดวก สามารถช่วยในการคัดเลือกกลุ่มคำที่น่าสงสัยว่าจะเป็นคำปรากฏร่วมออกมาให้เราได้

ในบทนี้จึงจะได้กล่าวถึงกลวิธีต่างๆ ในทางสถิติที่มีผู้ศึกษาและนำเสนอไว้ในสำหรับใช้ช่วยในการตัดสินหาว่าคำใดที่เป็นคำปรากฏร่วมกันได้บ้าง[1] คำที่เป็นคำปรากฏร่วมนั้นสังเกตุได้จากการที่พบปรากฏอยู่ติดกันเสมอ เช่น *disk*

---

[1] กลวิธีต่างๆที่ยกมานี้ เรียบเรียงจาก Oakes (1998) และ Manning and Schutze (1999)

drive, make up แต่ก็ไม่จำเป็นว่าคำที่เป็นคำปรากฏร่วมนั้นจะต้องปรากฏอยู่ติดกัน อาจจะมีคำอื่นๆมาคั่นระหว่างกลางได้ เช่น *throw ….. out*

4.1 การใช้ความถี่

ความคิดแรกในการสังเกตุหาคำที่เป็นคำปรากฏร่วม เราอาจทำได้โดยดูจากความถี่ของ bigram คือดูคำสองคำที่อยู่ติดกัน คู่คำที่ปรากฏร่วมกันบ่อย ก็น่าจะเป็นคำปรากฏร่วมกัน แต่ก็จะพบปัญหา เช่น จะพบว่าคู่คำอย่าง *of the, in the* จะปรากฏมากทั้งๆที่ไม่ใช่คำปรากฏร่วม ที่เป็นเช่นนี้ เพราะคำพวกนี้มีปรากฏใช้บ่อยจึงมีโอกาสที่จะปรากฏคู่กันสูงตามไปด้วย ดังนั้น เราจึงไม่สามารถดูคำปรากฏร่วมโดยพิจารณาจากความถี่ดิบเพียงอย่างเดียวได้ Justeson and Katz (1995) เสนอวิธีง่ายๆเพื่อแก้ปัญหานี้ โดยให้เลือกเฉพาะคู่คำที่หมวดคำของคำทั้งสองคำสามารถเป็นคำปรากฏร่วมกันได้ เช่น Adj-N, N-N, Adj-Adj-N เป็นต้น ซึ่งก็พบว่าผลที่ได้นี้มีความถูกต้องพอสมควร ความคิดของ Juteson and Katz นี้เป็นตัวอย่างของการแก้ปัญหาโดยใช้วิธีการทางสถิติแบบง่ายๆกับความรู้พื้นฐานทางภาษาศาสตร์ แต่วิธีการใช้ความถี่นี้ ก็ไม่สามารถชี้ให้เห็นคำที่ปรากฏร่วมกันตลอดเวลาแต่มีความถี่ในการเกิดไม่สูงมากนักได้

4.2 การใช้ค่าเฉลี่ยและความแปรปรวน

วิธีการใช้ความถี่ของ bigram จะใช้ได้สำหรับคำปรากฏร่วมที่มีลักษณะเป็นวลีตายตัว (fixed phrase) แต่ใช้ไม่ได้กับคำปรากฏร่วมที่มีคำแทรกอยู่ระหว่างกลางได้ เช่น ในตัวอย่าง *knock* กับ *door* ข้างล่างนี้

She knocked on his door

They knocked at the door

100 women knocked on Donaldson's door

a man knocked on the metal front door

(Manning and Schutze 1999: 147)

จะเห็นว่าคำที่ปรากฏอยู่ระหว่าง *knock* กับ *door* มีได้หลากหลายและมีจำนวนคำแตกต่างกัน ซึ่งจากลักษณะนี้ อาจนำแนวคิดเรื่องค่าเฉลี่ย (mean) และค่าความแปรปรวน (variance) หรือส่วนเบี่ยงเบนมาตรฐาน (SD) เพื่อดูการเกิด (distribution) ของสองคำนี้ โดยที่ค่าเฉลี่ยในที่นี้ คือ ค่าเฉลี่ยของระยะห่างระหว่างคำทั้งสอง ในตัวอย่างของ *knock* กับ *door* ข้างบนนั้น ค่าเฉลี่ยของระยะห่างคือ (3+3+5+5) / 4 เท่ากับ 3 (นับ Donalson's เป็น 3 คำ) ส่วนค่าความแปรปรวนคำนวณได้จากสูตร

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

ซึ่งในที่นี้จะได้ค่าส่วนเบี่ยงเบนมาตรฐาน

$$\sigma = \sqrt{\frac{1}{3}((3-4)^2 + (3-4)^2 + (5-4)^2 + (5-4)^2)} \approx 1.15$$

ค่าส่วนเบี่ยงเบนมาตรฐานนี้เป็นค่าที่บอกว่าระยะห่างระหว่างคำทั้งสองโดยทั่วไปมีความแตกต่างกันมากน้อยหรือไม่ ถ้ามีค่าเป็น 0 แสดงว่าคำทั้งสองนี้จะปรากฏในระยะห่างที่เท่ากันเสมอ ดังนั้น คำคู่ที่มีค่าเบี่ยงเบนมาตรฐานนี้ต่ำจึงน่าสงสัยว่าอาจจะเป็นคำที่เป็นคำปรากฏร่วมกันได้ ส่วนคู่คำที่มีค่าเบี่ยงเบนมาตรฐานสูงแสดงว่าไม่น่าจะใช่คำที่เป็นคำปรากฏร่วม ตัวอย่างในตารางข้างล่างแสดงผลจากการวิเคราะห์คลังข้อมูลของหนังสือพิมพ์ New York Times ข้อมูลที่แสดงแยกออกเป็นสามกลุ่ม กลุ่มแรกคือกลุ่มซึ่งมีค่า SD ต่ำคือมีค่าใกล้ศูนย์ เป็นกลุ่มที่แสดงถึงคำที่น่าจะเป็นคำปรากฏร่วมกัน กลุ่มที่สองเป็นกลุ่มที่มีค่า SD สูงซึ่งแสดงถึงคำที่ไม่มีลักษณะของการเป็นคำปรากฏร่วม ส่วนกลุ่มสุดท้ายเป็นกลุ่มซึ่งอยู่ระหว่างกลาง ยังไม่สามารถบอกได้ว่าเป็นคำปรากฏร่วมหรือไม่

SD	Mean	Count	Word1	Word2
0.43	0.97	11657	New	York
0.48	1.83	24	Previous	Games
0.15	2.98	46	Minus	Points
0.49	3.87	131	Hundreds	Dollars
4.03	0.44	36	Editorial	Atlanta
4.03	0.00	78	Ring	New
3.96	0.19	119	Point	Hundredth
3.96	0.29	106	Subscribers	By
1.07	1.45	80	Strong	Support
1.13	2.57	7	Powerful	Organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	Said

ตาราง 12.1 การหาคำปรากฏร่วมโดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

(Manning and Schutze 1999: 151)

Smadja (1993) ใช้วิธีการนี้เพื่อหาคำที่เป็นศัพท์เฉพาะสาขา (terminology term) ซึ่งก็ได้ผลค่อนข้างดีคือถูกต้องกว่า 80% แต่อย่างไรก็ดีคำปรากฏร่วมแบบที่ได้นี้ เช่น *knock - door* อาจไม่ใช่คำปรากฏร่วมในความหมายที่นักภาษาศาสตร์ต้องการ

4.3. การทดสอบสมมติฐาน

การทดสอบสมมติฐานเป็นวิธีการทางสถิติแบบที่ใช้กันทั่วไป ในที่นี้ถูกนำมาใช้เพื่อจะหาว่า การที่คำสองคำปรากฏร่วมกันนั้น ไม่ได้เป็นการปรากฏร่วมกันโดยบังเอิญ (by chance) สมมติฐานต้น (null hypothesis) คือให้คิดว่าการที่คำทั้งสองมาปรากฏร่วมกันนั้นเป็นเพราะเหตุบังเอิญ และคำสองคำนั้นไม่มีความสัมพันธ์ต่อกัน จากนั้นหาทางคำนวณค่าทางสถิติเพื่อดูว่าเราสามารถล้มล้างสมมติฐานต้น (reject null hypothesis) ที่ว่านี้ได้หรือไม่ จุดประสงค์ คือ เพื่อจะยืนยันว่าคำคู่นั้นอาจจะเป็นคำปรากฏร่วม เพราะค่าทางสถิติแสดงว่าคำคู่นั้นไม่ได้ปรากฏร่วมกันโดยเหตุบังเอิญ (เพราะสามารถล้มสมมติฐานต้นได้)

4.3.1 การทดสอบ t-test

t-test เป็นสถิติพื้นฐานที่ปกติจะใช้เพื่อทดสอบว่ากลุ่มตัวอย่างที่มีอยู่นั้นมาจากประชากร (population) เดียวกันหรือไม่โดยการเปรียบเทียบค่าเฉลี่ย (mean) และค่าความแปรปรวน (variance) ของกลุ่มตัวอย่างกับของกลุ่มประชากร สมมติฐานต้น (null hypothesis) คือกลุ่มตัวอย่างนั้นมาจากกลุ่มประชากรเดียวกันกับที่ศึกษา ค่า t-test ที่ใช้ในการคำนวณ คือ

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

การประยุกต์ใช้ t-test กับเรื่องคำปรากฏร่วม เช่น การหาว่า *new company* เป็นคำปรากฏร่วมหรือไม่ อาจทำได้โดยตั้งสมมติฐานต้นว่าสองคำนี้เป็นอิสระต่อกัน ดังนั้นโอกาสที่จะพบ *new company* หรือ P(new company) จะเท่ากับโอกาสที่จะพบ *new* หรือ P(new) คูณด้วยโอกาสที่จะพบ *company* หรือ P(company) ถ้าคลังข้อมูลมีขนาด 14,307,668 คำ มีคำว่า *new* ปรากฏอยู่ 15,828 ครั้ง คำว่า *company* ปรากฏ 4,675 ครั้ง ดังนั้น หากเป็นไปตามสมมติฐานต้น ค่าความน่าจะเป็นที่จะพบ *new* ปรากฏหน้า *company* คือ P(new company) = 15,828/14,307,668 x 4,675/14,307,668 = 3.615×10^{-7} หากคิดว่ามีกระบวนการที่จะสร้างตัวเลข 1 เมื่อพบ bigram “new company” และสร้างตัวเลขค่า 0 เมื่อพบ bigram อื่นๆ ค่าเฉลี่ยของกระบวนการนี้ก็จะเท่ากับ 3.615×10^{-7} ซึ่งถือได้ว่าเป็นค่าเฉลี่ยของประชากร μ แต่จากข้อมูลพบว่า *new* ปรากฏร่วมกับ *company* 8 ครั้ง P(new company) จริงนั้นเป็น 8/14,307,668 = 5.59110×10^{-9} ซึ่งถือได้ว่าเป็นค่าเฉลี่ย x-bar ส่วนค่าความแปรปรวนนั้นสามารถคำนวณได้ว่าเท่ากับ p * (1-p) ซึ่งประมาณว่าเท่ากับ p ได้ในกรณีของ

bigram เนื่องจากค่า p มีค่าต่ำมาก ดังนั้นจึงสามารถคำนวณค่า t-test ในตัวอย่างนี้ได้เท่ากับ

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

ค่า t ที่คำนวณได้นี้ไม่ถึง 2.576 ซึ่งเป็นค่าตัดสิน (critical value) สำหรับนัยยะสำคัญที่ 0.005 ดังนั้น เราจึงไม่สามารถล้มล้างสมมติฐานต้น *new* กับ *company* จึงเป็นอิสระจากกันได้ (หากเราสามารถล้มสมมติฐานที่ว่าค่าเฉลี่ยของกลุ่มตัวอย่างนี้มาจากประชากรเดียวกันกับ μ ซึ่งบอกว่าว่า *new* กับ *company* เป็นอิสระจากกัน เราก็สามารถสรุปได้ว่า *new* กับ *company* นั้นน่าสงสัยว่าจะเป็นคำปรากฏร่วมได้)

นอกจากการใช้ t-test เพื่อคำนวณว่าคำสองคำนั้นปรากฏร่วมกันอย่างมีนัยยะสำคัญทางสถิติ Church and Mercey (1993) ได้นำเอา t-test มาใช้เพื่อช่วยเปรียบเทียบว่าในคำสองคำที่กำหนด คำที่เป็นคำปรากฏร่วมกับคำทั้งสองนี้แตกต่างกันอย่างไร เช่น เมื่อเปรียบเทียบคำสองคำที่น่าสนใจ เช่น *strong* กับ *powerful* สิ่งที่เราสนใจอยากรู้ คือ ทั้งสองคำนี้มีความหมายแตกต่างกันอย่างไร ซึ่งหากเราสามารถหารายการคำที่เป็นคำปรากฏร่วมของ *strong* และรายการคำที่เป็นคำปรากฏร่วมของ *powerful* ก็จะช่วยให้เราเห็นความแตกต่างทางความหมายของคำทั้งสองได้ วิธีการที่ Church and Mercey ใช้คือคำนวณ t-test เพื่อเปรียบเทียบความต่างของคำใดใดกับคำทั้งสองนี้ โดยใช้สูตร t-test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(เนื่องจากสมมติฐานต้นคือความต่างของค่าเฉลี่ยระหว่างสองคำนี้เป็น 0 ทำให้ $\bar{x} - \mu = \bar{x} = \frac{1}{N}\sum(x_1 - x_2) = \overline{x_1} - \overline{x_2}$ ส่วนค่าความแปรปรวนนั้นใช้เป็นผลรวมของความแปรปรวนของประชากรทั้งสองกลุ่ม เนื่องจากค่าความแปรปรวนของความต่างของตัวแปรสองตัวจะเท่ากับผลรวมของความแปรปรวนของทั้งสองตัวแปร) และเนื่องจากค่าความแปรปรวนเท่ากับ p * (1-p) ซึ่งประมาณว่าเท่ากับ p ได้เพราะค่า p มีค่าน้อยมาก จึงสามารถคำนวณค่า t ของคำใดใดที่ปรากฏร่วมกับ *strong* และ *powerful* ได้ดังนี้

$$t \approx \frac{P(powerful\ w) - P(strong\ w)}{\sqrt{\frac{P(powerful\ w) + P(strong\ w)}{N}}} = \frac{\frac{C(powerful\ w)}{N} - \frac{C(strong\ w)}{N}}{\sqrt{\frac{C(powerful\ w) + C(strong\ w)}{N^2}}}$$

$$= \frac{C(powerful\ w) - C(strong\ w)}{\sqrt{C(powerful\ w) + C(strong\ w)}}$$

ตาราง 12.2 เป็นตัวอย่างที่ได้จากการคำนวณค่า t-score ของคำที่ปรากฏร่วมกับ *strong, powerful* จากสูตรดังกล่าว โดยเลือกคำที่ให้ค่า t สูงกว่า 2.576 ซึ่งแสดงนัยยะสำคัญที่ .005 ที่จะสามารถล้มสมมติฐานต้นที่ว่าไม่มีความแตกต่างในการปรากฏร่วมของคำนั้นกับ *strong* และ *powerful* หรือกล่าวอีกนัยหนึ่งว่าคำที่ให้ค่า t-score สูงกว่าค่า 2.576 นี้เป็นคำที่ชี้ให้เห็นความแตกต่างระหว่าง *strong* และ *powerful*

t	C(w)	C(strong w)	C(powerful w)	word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition
3.9000	802	18	1	showing
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

ตาราง 12.2 การหาคำปรากฏร่วมโดยใช้ t-test

(Manning and Schutze 1999: 157)

4.3.2 การทดสอบ Pearson's chi-square

การทดสอบโดยใช้ t-test มีข้อเสียประการหนึ่ง คือ เป็นการทดสอบที่ตั้งอยู่บนข้อสันนิษฐานว่าข้อมูลมีการกระจายตัวแบบสุ่ม (randomly distribute) ซึ่งในความเป็นจริง ภาษาไม่ได้มีลักษณะเป็นเช่นนั้น (Church and Mercy 1993) สถิติที่ใช้ในการทดสอบสมมติฐานที่ไม่จำเป็นต้องตั้งอยู่บนข้อสันนิษฐานว่าข้อมูลมีการกระจายตัวอย่างสม่ำเสมอ คือการทดสอบค่า chi-square

Chi-square เป็นค่าที่ใช้ทดสอบความแตกต่างระหว่างความถี่ของการปรากฏที่สังเกตุได้กับความถี่ของการปรากฏที่คาดว่าอาจจะเกิดได้โดยบังเอิญ

เพื่อดูว่าความแตกต่างของค่าทั้งสองนี้มีนัยยะสำคัญทางสถิติหรือไม่ ในที่นี้สามารถนำมาทดสอบว่าคำสองคำนั้นเป็นคำปรากฏร่วมหรือไม่ โดยการเปรียบเทียบความถี่ของการปรากฏของทั้งสองคำ ตัวอย่างเช่น ถ้าต้องการตรวจสอบว่า *new* กับ *companies* เป็นคำปรากฏร่วมกันหรือไม่ ซึ่งสมมติว่าความถี่ของคำทั้งสองเป็นดังในตาราง ดังนี้

	W1 = new	W1 ≠ new
W2 = companies	8 new companies	4667 e.g. old companies
W2 ≠ companies	15820 e.g. new machines	14287181 e.g. old machines

ตาราง 12.3 การเกิดร่วมกันของ new, companies, และคำอื่น

(Manning and Schutze 1999: 158)

จากคลังข้อมูลภาษา เราพบ *new* เกิดร่วมกับ *companies* 8 ครั้ง และคำอื่นๆเกิดกับ *companies* 4,667 ครั้ง *new* เกิดกับคำอื่นๆ 15,820 ครั้ง และคำอื่นๆที่ไม่ใช่ *new* กับคำที่ไม่ใช่ *companies* เกิด 14,287,181 ครั้ง จากนั้นคำนวณหาค่า chi-square โดยใช้สูตร $\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

โดยที่ $O_{i,j}$ คือค่าที่พบในแถวที่ i คอลัมน์ที่ j $E_{i,j}$ คือค่าที่คาดไว้ในแถวที่ i คอลัมน์ที่ j ซึ่งค่าที่คาดไว้ของแต่ละช่องจะเท่ากับ (row total X column total) / grand total ในที่นี้ ค่าที่คาดไว้ของช่อง (1,1) คือ (8+4,667) * (8+15,820) / (8+4,667+15,820+14,287,181) เท่ากับ 5.2 (หมายความว่าโอกาสที่จะพบ *new* กับ *companies* เกิดร่วมกันถึงแม้ว่าจะไม่ได้เป็นคำปรากฏร่วมกันจะเท่ากับ 5.2 ครั้ง) ในตัวอย่างนี้ค่าที่คาดเอาไว้ของแต่ละช่องจะเป็นดังนี้

$E_{1,1}$ = (8+4,667) * (8+15,820) / (8+4,667+15,820+14,287,181) = 5.17

$E_{1,2}$ = (8+4,667) * (4,667+14,287,181) / (8+4,667+15,820+14,287,181) = 4,669.83

$E_{2,1}$ = (15,820+14,287,181) * (8+15,820) / (8+4,667+15,820+14,287,181) = 15,822.83

$E_{2,2}$ = (15,820+14,287,181) * (4,667+14,287,181) / (8+4,667+15,820+14,287,181) = 14,287,178.17

ค่า chi-square ในที่นี้จะเท่ากับ $(8-5.17)^2/5.17 + (4,667-4,669.83)^2/4,669.83 + (15,820-15,822.83)^2/15,822.83 + (14,287,181-14,287,178.17)^2/14,287,178.17 = 1.55 + 0 + 0 + 0$

ค่า chi-square ที่มีนัยยะสำคัญ 0.05 จะอยู่ที่ 3.84 ดังนั้น ในที่นี้จึงไม่สามารถลบล้างสมมติฐานต้นที่ว่าการปรากฏร่วมกันของ *new* กับ *companies* นั้นเป็นไปโดยบังเอิญ *new companies* จึงไม่ใช่คู่ที่น่าสงสัยว่าจะเป็นคำปรากฏร่วมกัน

4.3.3 การทดสอบ Berry-Rogghe's z-score

Berry-Rogghe (1973) ใช้วิธีหารายการของคำที่มีแนวโน้มว่าจะเป็นคำปรากฏร่วมของแต่ละคำ โดยใช้วิธีดูความน่าจะเป็นที่คำๆหนึ่ง จะเกิดร่วมกับคำอื่นๆ ภายในขอบเขต (span) ที่กำหนด ถ้ากำหนดให้

Z เป็นจำนวนคำทั้งหมดภายในเอกสาร

A เป็นคำที่ต้องการศึกษาซึ่งปรากฏ Fn ครั้งในเอกสาร

B เป็นคำที่เป็นคำปรากฏร่วมของ A ซึ่งปรากฏ Fc ครั้งในเอกสาร

K เป็นจำนวนครั้งของการเกิดร่วมกันของ B และ A

S เป็นขนาดของขอบเขต (span) หรือคือจำนวนคำที่อยู่ข้างใดข้างหนึ่งของคำที่ต้องการ

ในที่นี้ A คือคำที่ต้องการศึกษา สิ่งที่ต้องการศึกษาคือ หาว่าคำอะไรบ้างที่จะเป็นคำปรากฏร่วมของ A วิธีคิดเป็นดังนี้ ในขั้นแรกให้คำนวณความน่าจะเป็นที่จะพบ B ปรากฏร่วมกับ A เป็นจำนวน K ครั้ง โดยสมมติว่า B จะปรากฏแบบสุ่ม (random) ที่ใดก็ได้ ค่าที่ได้นี้ เรียกว่าค่าความน่าจะเป็นที่คาดว่าจะพบ B ปรากฏร่วมกับ A จากนั้นจึงหาความแตกต่างระหว่างค่าที่คาดไว้ (expected number) กับค่าที่พบจริง (observed number)

ก่อนอื่น คำนวณค่าความน่าจะเป็นที่จะพบ B ปรากฏในตำแหน่งใดใดที่ A ไม่ปรากฏอยู่ ซึ่งจะเท่ากับ

p = Fc / (Z - Fn)

จากนั้น คำนวณจำนวนครั้งที่จะพบ B ปรากฏร่วมกับ A ภายในขอบเขต S ซึ่งได้เท่ากับ E = p * Fn * S

จากนั้นคำนวณว่า ค่าที่พบจริงๆ (observed) กับค่าที่คำนวณได้ (expected) มีความแตกต่างอย่างมีนัยยะสำคัญทางสถิติหรือไม่ ซึ่งสามารถคำนวณได้จาก z-score ดังนี้

$z = (K - E) / \sqrt{Eq}$ โดยที่ q = 1-p

ถ้าจะให้มีนัยยะสำคัญที่ระดับ 1 % ค่า z-score จะต้องมากกว่า 2.576 และเมื่อรวมค่า z-score ของคำทุกคำที่คำนวณเทียบกับคำ A ผลที่ได้ควรจะมีค่าเป็น 0 ข้อจำกัดของวิธีทดสอบแบบนี้ คือ ไม่ยอมให้คำใดใดเป็นคำปรากฏร่วมของตัวเอง Berry-Roggle ใช้วิธีนี้คำนวณหาคำปรากฏร่วมของคำว่า *house* ในคลังข้อมูลที่ประกอบด้วยหนังสือ A Christmas Carol, Each in his own Wildernes (แต่งโดย Doris Lessing) และ Everything in the Garden (แต่งโดย Giles Cooper) และทดลองกำหนดค่าขอบเขต (span) ต่างๆ จาก 3 ถึง 6 เพื่อหาค่าขอบเขตที่ดีที่สุด ซึ่งก็พบว่าเมื่อใช้ค่าขอบเขตเป็น 3 จะได้คำปรากฏร่วมของ *house* เป็น *sold,*

commons, decorate, this, empty, buying, painting, opposite, loves, outside, lived, family, remember, full, my, into, the, has แต่เมื่อตั้งขอบเขตให้มีค่าเป็น 6 คำที่เป็นคำปรากฏร่วมของ *house* ที่ได้ คือ *sold, commons, decorate, fronts, cracks, this, empty, buying, painting, opposite, loves, entered, black, near, outside, remember, lived, rooms, God, stop, garden, flat, every, big, my, into, family, Bernard, whole* นั่นคือ การเพิ่มค่าขอบเขตจะทำให้สามารถขจัดคำปรากฏร่วมที่ไม่ต้องการ เช่น *the* และ *has* ได้ และทำให้พบคำปรากฏร่วมคำอื่นๆ เช่น *rooms, God, garden* หรือ *flat* แต่ก็ทำให้ได้คำปรากฏร่วมบางคำที่ไม่ต้องการออกมาด้วย เช่น *Bernard* Berry-Rogghe พบว่าค่าขอบเขตที่ควรจะใช้โดยทั่วไปคือ 4 ยกเว้นกรณีที่ต้องการหาคำปรากฏร่วมของคำคุณศัพท์ซึ่งควรจะใช้ค่าขอบเขตเป็น 2

ในปี 1974 Berry-Rogghe ใช้วิธีการเดียวกันศึกษาเรื่องกริยาบุพบท (phrasal verb) หรือกริยาที่ปรากฏร่วมกับคำที่เป็น particle ตัวอย่างเช่น *look after, give in* ซึ่งกริยาบุพบทจะทำหน้าที่เหมือนเป็นคำๆ เดียว Berry-Rogghe ศึกษาโดยเอาคำ particle เป็นตัวตั้ง จากนั้นหาคำกริยาที่ปรากฏหน้า particle นั้น วิธีนี้จะสะดวกกว่าการเอาคำกริยาเป็นตัวตั้งแล้วหา particle ที่ปรากฏร่วมกับคำกริยานั้นเพราะว่าจำนวนคำที่เป็น particle มีน้อยกว่าจำนวนคำกริยา ตัวอย่างเช่น เมื่อหาคำปรากฏร่วมของคำกริยาที่เกิดหน้า in ก็จะได้รายการของคำปรากฏร่วม ดังนี้ *interested, versed, lived, believe, found, live, ride, living, dropped, appeared, travelled, die, sat, died, interest, life, rode, stood, walk, find, house, arrived, came* ส่วนการที่จะตัดสินว่า กริยาตัวไหน เมื่อเกิดร่วมกับ *in* แล้วเป็น idiomatic phrasal verb อาศัยหลักที่ว่า ความหมายของ idiomatic phrasal verb ควรจะต่างจากความหมายของส่วนประกอบย่อย ซึ่ง

สามารถสังเกตความต่างนี้ได้จากคำปรากฏร่วมที่ต่างกัน ตัวอย่างเช่น คำปรากฏร่วมของ *hot dog* จะเป็น *eat, mustard, stall* ในขณะที่คำปรากฏร่วมของ *hot* จะเป็น *weather, air, water* และคำปรากฏร่วมของ *dog* จะเป็น *bark, tail* *hot dog* จึงมีลักษณะเป็นหน่วยคำเดียวมากกว่าจะเป็นสองคำ Berry-Rogghe (1974: 21-22) จึงกล่าวว่า

> The combination of verb + particle are to be constituting a single lexical item when they contract different collocational relations from those of the particle as a separate entity.

ดังนั้น Berry-Rogghe จึงหาคำปรากฏร่วมของคำที่เกิดทางขวาของ *in* (โดยนับเฉพาะ *in* ที่ตามหลัง punctuation mark เพื่อเลี่ยงผลกระทบจากคำปรากฏร่วมที่อาจจะมาจากคำกริยาที่อยู่หน้า *in*) ผลที่ได้จะได้รายการของคำที่เป็นคำปรากฏร่วมทางขวาของ *in* เดี่ยวๆ จากนั้น นำคำที่ได้จากรายการแรกที่ได้มาก่อน คือ รายการคำกริยาที่สงสัยว่าจะรวมกับ *in* เป็นกริยาบุพบท (phrasal verb) เช่น *interested in, versed in, live in* เป็นต้น มาหาคำปรากฏร่วมทางขวาของ phrasal verb เหล่านี้ด้วย แล้วคำนวณหาค่า R score โดยที่กำหนดให้ R = a/b โดยที่ a เป็นจำนวนคำที่เป็นคำปรากฏร่วมของทั้ง *in* และกริยาบุพบท (V + in) ส่วน b เป็นจำนวนคำที่เป็นคำปรากฏร่วมของ phrasal verb ตัวอย่าง เช่น *versed in* มีคำที่เป็นคำปรากฏร่วมที่สำคัญอยู่ 3 ตัว คือ *politics, history, Greek* แต่ไม่มีคำไหนเลยที่เป็นคำปรากฏร่วมของ *in* ดังนั้น R = 0/3 = 0 ส่วน *live in* มีคำที่เป็นคำปรากฏร่วมที่สำคัญอยู่ 11 คำ คือ *hut, house, *town, *country, *London, *room, *world, *place, family, happiness, ignorance* ในจำนวนนี้มีอยู่ 6 คำที่เป็นคำปรากฏร่วมของ *in* ด้วย คือคำที่มีเครื่องหมาย * ใส่ไว้ข้างหน้า ดังนั้น R ของ *live in* = 6/11 = 0.54 ยิ่ง

ตัวเลขที่ได้ต่ำเท่าไร กริยาบุพบทที่สงสัยนั้นก็มีแนวโน้มที่จะเป็น idiomatic มากขึ้น

วิธีทดสอบแบบ Berry-Rogghe's z-score เพื่อหาคำปรากฏร่วมนั้นใช้อยู่ในโปรแกรมคอนคอร์แดนซ์ TACT และในโปรแกรม SARA ซึ่งเป็นโปรแกรมที่ใช้ร่วมกับคลังข้อมูล BNC

4.3.4 การทดสอบค่า Likelihood ratio

การทดสอบด้วย Dunning's likelihood ratio นี้เหมาะสมกว่าการใช้ค่า chi-square ค่าที่ได้นี้จะบอกถึงปริมาณว่าสมมติฐานหนึ่งมีความคล้ายคลึงกับอีกสมมติฐานหนึ่งเพียงใด ในการหาคำปรากฏร่วม จะมีสมมติฐานที่จะเปรียบเทียบสองสมมติฐานดังนี้

Hypothesis 1 คือ P(w2|w1) = p = P(w2| ¬w1) (w2 เป็นอิสระจาก w1)

Hypothesis 2 คือ P(w2|w1) = p1 ≠ p2 = P(w2|¬w1) (w2 นั้นขึ้นอยู่กับ w1)

ซึ่ง p = C(w2)/N, p1 = C(w1 w2)/C(w1), และ p2 = (C(w2) – C(w1 w2)) / (N-C(w1)) หรือเขียนแบบสั้นๆว่า p = c2/N, p1 = c12/c1 และ p2 = (c2 – c12) / (N-c1)

และหากเราถือเอาว่าข้อมูลมีการกระจายตัวแบบ binomial distribution

$$b(k,n,k) = \binom{n}{k} x^k (1-x)^{(n-k)}$$

ค่า Likelihood L(H1) = b(c12, c1,p) b(c2-c12, N-c1, p) และค่า L(H2) = b(c12, c1,p1) b(c2-c12, N-c1, p2)

จากนั้นคำนวณค่า log ของอัตราส่วน likelihood ดังนี้

$$\log \lambda = \log \frac{L(H1)}{L(H2)} = \log \frac{b(c12, c1, p)b(c1 - c12, N - c1, p)}{b(c12, c1, p1)b(c1 - c12, N - c1, p2)}$$
$$= \log L(c12, c1, p) + \log L(c2 - c12, N - c1, p) - \log L(c12, c1, p1) -$$
$$\log L(c2 - c12, N - c1, p2)$$
$$where\ L(k, n, x) = x^k (1 - x)^{n-k}$$

ค่า log likelihood นี้สามารถนำมาตีความได้โดยตรง หรือจะนำทดสอบสมมติฐานแบบเดียวกับค่า chi-square ก็ได้ โดยคูณค่าที่ได้ด้วย –2 เป็น –2 log λ จากนั้นเปิดดู critical value ของตาราง chi-square ค่านัยยะสำคัญที่ 0.005 (โดยมี degree of freedom = 1) ต้องมากกว่า 7.88 จึงจะล้มสมมติฐานต้นว่าคำทั้งสองนั้นเป็นอิสระจากกันได้

4.4 การใช้ค่า Mutual Information

ค่า mutual information เป็นค่าที่วัดจากการเปรียบเทียบความน่าจะเป็นของการพบคำสองคำเกิดด้วยกัน หารด้วยความน่าจะเป็นที่จะพบคำทั้งสองโดยอิสระ ตามสูตรต่อไปนี้

$$I(x', y') = \log_2 \frac{P(x'\, y')}{P(x')P(y')}$$

ถ้าคำทั้งสองเป็นคำปรากฏร่วม ค่าความน่าจะเป็นที่จะพบคำทั้งสองปรากฏร่วมกัน (P(X' Y')) จะสูงกว่าค่าความน่าจะเป็นที่จะพบคำทั้งสองปรากฏร่วมกันโดยบังเอิญ (P(X') * P(Y')) ผลลัพธ์ที่ได้จะบอกค่าปริมาณของ information เกี่ยวกับการปรากฏของคำ X ที่ตำแหน่ง i และคำ Y ที่ตำแหน่ง i+1 ถ้าค่าที่ได้ยิ่งมีค่าสูงก็ยิ่งแสดงว่าคำทั้งสองนั้นก็มีแนวโน้มที่จะเป็นคำปรากฏร่วม ตัวอย่าง เช่น ถ้าคำ *Bette* ปรากฏ 41 ครั้ง *Midler* ปรากฏ 27 ครั้ง และ *Bette Midler* ปรากฏ 20 ครั้ง จากข้อมูลทั้งหมด 14,307,668 คำ ค่า I(Bette,Midler) จะเท่ากับ $\log_2$ (20/14,307,668) / (41/14,307,668 * 27/14,307,668) จะได้

ประมาณ 17.98 ด้วยวิธีนี้ เราสามารถเรียงลำดับของคำทุกคู่ตามค่า mutual information จากมากไปน้อย คู่ที่ได้อยู่ส่วนบนๆ ของรายการคือคู่ที่น่าสงสัยว่าจะเป็นคำปรากฏร่วม

ตารางข้างล่างเป็นตัวอย่างของการคำนวณค่า MI จากคลังข้อมูล 1988 Associated Press Corpus (จำนวนคำ = 44.3 ล้านคำ) (Church et al. 1991: 121) ซึ่งคำว่า *strong* ปรากฏ 7,809 ครั้ง คำว่า *northerly* ปรากฏ 28 ครั้ง

I(x;y)	fxy	fx	Fy	x	y
10.47	7	7809	28	strong	northerly
9.76	23	7809	151	strong	showings
9.30	7	7809	63	strong	believer
9.22	14	7809	133	strong	second-place
9.17	6	7809	59	strong	runup
9.04	10	7809	108	strong	currents
8.85	62	7809	762	strong	supporter
8.84	8	7809	99	strong	proponent
8.68	15	7809	208	strong	thunderstorm
8.45	7	7809	114	strong	odor

ตาราง 12.4 คำปรากฏร่วมของ strong เรียงตามค่า MI

I(x;y)	fxy	fx	fy	x	y
8.66	7	1984	388	powerful	legacy
8.58	7	1984	410	powerful	tool
8.35	8	1984	548	powerful	storms
8.32	31	1984	2169	powerful	minority
8.14	9	1984	714	powerful	neighbor
7.98	9	1984	794	powerful	Tamil
7.93	8	1984	734	powerful	symbol
7.74	32	1984	3336	powerful	figure
7.54	10	1984	1204	powerful	weapon
7.47	24	1984	3029	powerful	post

ตาราง 12.5 คำปรากฏร่วมของ powerful เรียงตามค่า MI

แต่ข้อเสียของการใช้ค่า MI คือ ตัวเลขที่ได้ขึ้นอยู่กับค่าความถี่ของการปรากฏ เช่น ถ้าพิจารณากรณีสุดโต่งที่ X จะปรากฏร่วมกับ Y ทุกครั้ง เราจะได้

$I(X,Y) = \log_2 P(X Y)/P(X)P(Y) = \log_2 P(X)/P(X)P(Y) = \log_2 1/P(Y)$ (เพราะว่า P(X Y) จะเท่ากับ P(X))

ค่าที่ได้จะขึ้นอยู่กับ P(Y) ส่วนในกรณีที่การปรากฏของ X ไม่ขึ้นอยู่กับ Y เลย เราจะได้

$I(X,Y) = \log_2 P(X Y)/P(X)P(Y) = \log_2 P(X)P(Y)/P(X)P(Y) = \log_2 1 = 0$

ดังนั้น ถ้ามองในแง่นี้ ค่า MI ใช้บอกความเป็นความเป็นอิสระ (independence) ระหว่างคำได้ดี กล่าวคือค่ายิ่งเข้าใกล้ 0 มากเท่าใดคำสองคำนั้นยิ่งเป็นอิสระต่อกันมากขึ้น แต่ค่า MI ไม่สามารถบอกความผูกพันที่มีต่อกัน (dependence) เพราะค่าที่ได้ในกรณีของคำที่มีความผูกพันแบบสมบูรณ์ (perfect dependence) นั้นเปลี่ยนไปตามค่า P(Y) ไม่รู้ค่าสูงสุด (upper bound) ที่แน่นอน นอกจากนี้ ค่า MI ที่คำนวณได้จากคู่คำที่มีความถี่ในการเกิดต่ำ เช่น เกิดเพียงหนึ่งครั้งก็มีแนวโน้มจะค่าที่สูงกว่าค่า MI ที่คำนวณจากคู่คำที่มีความถี่ในการเกิดสูง ซึ่งไม่น่าจะเป็นสิ่งที่เราต้องการ ดังนั้น ค่า MI จึงไม่ใช่วิธีการที่ดีในการหาคำปรากฏร่วม บางคนจึงหลีกเลี่ยงปัญหาอันนี้โดยเลือกคำนวณค่า MI เฉพาะกับคำที่มีความถี่มากกว่าที่กำหนด เช่น 3 เป็นต้น หรือชดเชยค่าโดยนำเอาค่าความถี่มาคำนวณด้วยเป็น $C(X Y) * I(X,Y)$

4.5 การทดสอบแบบอื่น ๆ

Daille (1995) ได้ทดลองใช้เทคนิคแบบต่างๆ ในการคำนวณหาคำที่เป็นคำปรากฏร่วม ทั้งนี้หากเราพิจารณาคำที่ปรากฏร่วมกันเป็นคู่ ๆ ของ L_i และ L_j โดยที่

a เป็นความถี่ของคู่ที่ประกอบด้วย L_i และ L_j

b เป็นความถี่ของคู่ที่ประกอบด้วย L_i แต่ไม่มี L_j

c เป็นความถี่ของคู่ที่ประกอบด้วย L_j แต่ไม่มี L_i

d เป็นความถี่ของคู่ที่ไม่มีทั้ง Li และ Lj

	$W_1 = L_i$	$W_1 \neq L_i$
$W_2 = L_j$	a	c
$W_2 \neq L_j$	b	d

ดังนั้น N = a+b+c+d เป็นจำนวนของคู่คำ (bigram) ทั้งหมดในเอกสาร วิธีที่จะวัดหาระดับของความสัมพันธ์ระหว่างคำ 2 คำ สามารถวัดได้หลายๆ แบบ ดังเช่น

Simple matching coefficient (SMC) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$SMC = \frac{a+d}{a+b+c+d}$$

Kulczinsk coeffcient (KUC) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$KUC = \frac{a}{2}(\frac{1}{a+b} + \frac{1}{a+c})$$

Ochiai coefficient (OCH) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Frager and McGowan coefficient (FAG) ซึ่งให้ค่าตั้งแต่ - infinity ถึง 1

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{(a+b)}}$$

Yule coefficient (YUL) ซึ่งให้ค่าระหว่าง -1 ถึง 1

$$YUL = \frac{ad-bc}{ad+bc}$$

Phi-squared coefficient ซึ่งให้ค่าระหว่าง 0 ถึง infinity

$$\Phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

Cubic association ratio (MI3)

เนื่องจากค่า Mutual Information ให้ความสำคัญกับเหตุการณ์ที่มีความถี่ต่ำมากไป จึงมีผู้เสนอ cubic association ratio (MI3) เพื่อให้น้ำหนักกับเหตุการณ์ที่เกิดบ่อยมากขึ้น

$$MI3 = \log_2 \frac{a^3 N}{(a+b)(a+c)}$$

และ Log-likelihood coefficient (LL) ซึ่งเสนอโดย Dunning (1993)

LL = 2 x (a log a + b log b) + c log c) + d log d

-(a + b) log (a + b) - (a + c) log (a + c)

-(b + d) log (b + d) - (c + d) log (c + d)

+ (a + b + c + d) log (a + b + c + d)

โดยที่ log ในที่นี้เป็น log ฐาน e

Daille (1995) ทดสอบค่า coefficient เหล่านี้ เพื่อที่จะหาคู่คำที่เป็นศัพท์เทคนิค (technical term) โดยนำผลทั้งหมดที่ได้ไปเปรียบเทียบกับศัพท์ที่เป็นศัพท์เทคนิคจริงๆ หรือจากศัพท์ที่ผู้เชี่ยวชาญ 2 ใน 3 คนตัดสินว่าเป็นศัพท์เทคนิค Daille สรุปผลว่า วิธีทดสอบที่ได้ผลน่าพอใจ คือ log-likelihood, Frager and McGowan coefficient, และ cubic association ratio และในวิธีการเหล่านี้ log-likelihood เป็นวิธีการที่ดีที่สุด

4.6 การหาคำปรากฏร่วมมากกว่าสองคำ

วิธีการทางสถิติที่กล่าวมาข้างต้นช่วยในการหาคำปรากฏร่วมระหว่างคำสองคำซึ่งอาจปรากฏติดกันหรือไม่ก็ได้ แต่ยังไม่ครอบคลุมกรณีของคำปรากฏร่วมที่ประกอบด้วยคำมากกว่าสองคำ Kita et al. (1994) จึงเสนอแนวคิดที่ใช้ reduced cost ในการหาคำปรากฏร่วมของกลุ่มคำหลายคำ เช่น *in spite of* โดยที่ถ้าให้ a แทน *in spite* และ b แทน *in spite of* ค่า reduced cost, K(a), ที่

คำนวณจะเท่ากับ K(a) = (|a| - 1) x (f(a) –f(b)) โดยที่ |a| หมายถึงจำนวนคำของ a (ในตัวอย่างนี้มีค่าเป็น 2) ซึ่งเมื่อคำนวณแล้ว f(in spite) จะไม่ต่างจาก f(in spite of) มากนัก ค่าที่ได้จึงต่ำ ในขณะเดียวกัน ถ้าคำนวณ *in spite of the* ค่า f(in spite of) – f(in spite of the) จะมีค่าต่างกันมาก ค่าที่คำนวณได้จึงสูง วิธีการนี้จึงช่วยระบุได้ว่า *in spite of* น่าจะเป็นคำปรากฏร่วมได้

4.7 การหาคำปรากฏร่วมระหว่างสองภาษา

นอกจากการหาคำปรากฏร่วมภายในคลังข้อมูลภาษาเดียว (monolingual corpora) แล้ว เรายังสามารถหาคำปรากฏร่วมที่ปรากฏในคลังข้อมูลสองภาษา (bilingual corpora) ได้ด้วย ถ้าเรามีคลังข้อมูลสองภาษาที่มีการจับคู่ (align) ประโยคระหว่างภาษาทั้งสอง เราอาจใช้สถิติแบบที่ได้กล่าวมาแล้วเพื่อวัดระดับความสัมพันธ์ของคำปรากฏร่วม (collocation) ระหว่างภาษาได้ ซึ่งผลที่จะได้คือคำที่น่าจะเป็นคำแปลระหว่างสองภาษานั้น ทั้งนี้เพราะประโยคที่ถูกจับคู่ (align) ระหว่างสองภาษา คือ ประโยคที่เป็นประโยคแปลของอีกภาษาหนึ่ง ตัวอย่างเช่น Gaussier, Lange, and Meunier วัดระดับความสัมพันธ์ของคำปรากฏร่วมระหว่างคำภาษาอังกฤษและฝรั่งเศส โดยใช้ค่า mutual information ซึ่งคำนวณได้จาก

$$I(e, f) = \log_2 \frac{p(e, f)}{p(e)p(f)}$$

โดยที่ p(e,f) เป็นค่าความน่าจะเป็นที่จะพบคำ e และ f ในประโยคที่ถูกจับคู่กัน (aligned sentences) p(e) เป็นค่าความน่าจะเป็นที่จะพบคำ e ในประโยคภาษาอังกฤษ p(f) เป็นค่าความน่าจะเป็นที่จะพบคำ f ในประโยคภาษาฝรั่งเศส ตัวอย่าง เช่น สำหรับคำภาษาอังกฤษ prime มีคำภาษาฝรั่งเศสที่น่าจะเป็นคำแปลของคำนี้ คือ

sein	MI=5.63	bureau	MI=5.63	trudeau	MI=5.34
premier	MI=5.25	residence	MI=5.12	intention	MI=4.57
no	MI=4.53	session	MI=4.34		

จากนั้นตรวจสอบว่าคำเหล่านี้มีคำไหนบ้างที่มีค่า MI สูงกับคำภาษาอังกฤษคำอื่น แล้วให้ตัดคำเหล่านั้นออกไป ซึ่งจะเหลือคำ *premier* เพียงคำเดียว จึงถือเป็นคำที่เป็นคำปรากฏร่วมของ *prime* จากการศึกษา พบว่าสามารถหาคำแปลที่ถูกต้องของคำภาษาอังกฤษได้ 65% อีก 25% ไม่สามารถหาคำแปลภาษาฝรั่งเศสได้ (เนื่องจากไม่มีคำฝรั่งเศสที่แปลได้โดยตรง) และประมาณ 10% เลือกคำแปลมาผิด

นอกจาก การหาคำปรากฏร่วมของคำระหว่างภาษาแล้ว เราก็สามารถหาคำปรากฏร่วมของวลีระหว่างภาษาได้เหมือนกัน เช่น การหาคำปรากฏร่วมของศัพท์เทคนิค (technical term) ระหว่างภาษาอังกฤษและฝรั่งเศส เช่น ในงานของ Gaussier and Lange (1994) แต่รายละเอียดจะไม่กล่าวถึงในที่นี้

5. สรุป

บทนี้ได้นำเสนอวิธีการหาคำปรากฏร่วมโดยวิธีการทางสถิติแบบต่างๆ ซึ่งจะเป็นประโยชน์สำหรับการศึกษาภาษาโดยอาศัยคลังข้อมูลภาษาเป็นพื้นฐาน วิธีการต่างๆ เหล่านี้สามารถช่วยให้เราค้นหาคำปรากฏร่วมของคำต่างๆ ได้โดยอัตโนมัติ แน่นอนว่า ผลที่ได้ทุกรายการอาจไม่ใช่คำปรากฏร่วมแบบที่นักภาษาศาสตร์สนใจ แต่วิธีนี้ก็สามารถช่วยให้นักภาษาศาสตร์ทำงานได้สะดวกมากขึ้น เพราะเป็นการกลั่นกรองข้อมูลที่น่าสงสัยว่าจะเป็นคำปรากฏร่วมออกมาให้วิเคราะห์ ผู้สนใจสามารถดาวน์โหลดโปรแกรม Collocation Test เพื่อใช้ช่วยในการค้นหาคำปรากฏร่วมได้ที่ http://pioneer.chula.ac.th/~awirote/colloc/

บทที่ 13 การใช้เว็บเป็นคลังข้อมูล

1. บทนำ

ที่ผ่านมาเราได้เห็นว่าเราสามารถใช้ตัวบทที่ปรากฏในเว็บไซต์ต่างๆ เพื่อรวบรวมเป็นคลังข้อมูลภาษาแบบที่ต้องการ แต่ก็มีผู้ที่สนใจว่า ตัวเว็บเองสามารถที่จะใช้เป็นเสมือนคลังข้อมูลภาษาได้หรือไม่ หากเว็บเป็นคลังข้อมูลภาษาจะเป็นตัวแทนของภาษาประเภทไหน หากเว็บไม่ใช่คลังข้อมูลภาษาที่กำหนดได้ชัดเจน เราจะยังสามารถใช้ประโยชน์จากเว็บเพื่อการศึกษาเกี่ยวกับภาษาได้หรือไม่และใช้ได้อย่างไร และหากเราสามารถใช้ประโยชน์จากเว็บได้แล้ว ยังมีความจำเป็นที่จะต้องสร้างคลังข้อมูลภาษาขนาดใหญ่อย่าง BNC หรือ ANC อีกหรือไม่

2. ข้อมูลภาษาในเว็บ

ข้อมูลภาษาบนเว็บเป็นข้อมูลที่อาจมองได้ว่าเป็นคลังข้อมูลแบบสั่งสม (monitor corpus) เหมือนอย่าง Bank of English เพราะมีการเพิ่มเติมข้อมูลเข้าสะสมตลอดเวลา แต่ก็ไม่ใช่คลังข้อมูลภาษาแบบสั่งสม เพราะเราไม่รู้ขนาดที่ชัดเจนของข้อมูลว่ามีเท่าใด ทำได้เพียงประมาณการจำนวนหน้าเว็บอย่างเช่น Google ในปีค.ศ. 2002 ได้ทำดัชนีหน้าเว็บได้ 3 พันล้านหน้าเว็บ (Meyer et.al. 2003) ในปีค.ศ.2005 Gulli and Signorini ได้ศึกษาและประมาณหน้าเว็บบนอินเตอร์เน็ตว่ามีประมาณกว่า 11.5 พันล้านหน้าเว็บ และอัตราการขยายตัวก็มี

เพิ่มขึ้นทวีคูณทุกปี Google เองก็ได้รายงานเมื่อวันที่ 25 ก.ค. 2008 ว่า[1] ได้พบลิ้งค์หรือ URL ที่ชี้ไปยังเว็บที่ไม่ซ้ำกันได้ถึง 1 ล้านล้านเว็บ (1 trillion) เมื่อเทียบกับที่ Google เริ่มต้นทำดัชนีเว็บในปี 1998 ที่มีเพียง 26 ล้านเว็บและในปี 2000 ที่ได้มาถึง 1 พันล้านเว็บ อินเทอร์เน็ตจึงเป็นแหล่งข้อมูลภาษาที่มีขนาดใหญ่กว่าคลังข้อมูลภาษาใดๆ และมีขนาดที่ไม่จำกัด

อย่างไรก็ตาม ข้อมูลที่อยู่บนเว็บนั้นก็มีข้อมูลภาษาต่างๆทั่วโลก เมื่อเริ่มต้นข้อมูลส่วนใหญ่จะเป็นภาษาอังกฤษ แต่ภาษาอื่นๆก็เริ่มปรากฏขึ้นบนเว็บมากขึ้นเรื่อยๆ หากพิจารณาข้อมูลผู้ใช้อินเทอร์เน็ตจาก Internet World Stat ซึ่งเป็นเว็บไซต์รายงานจำนวนการใช้งานอินเทอร์เน็ตว่า เมื่อสิ้นปี ค.ศ.2009 มีผู้ใช้งานอินเทอร์เน็ตทั่วโลกกว่า 1,802 ล้านคน เปรียบเทียบกับในปีค.ศ.2000 ซึ่งมีเพียง 360 ล้านคน และจำนวนที่มากที่สุดนี้มาจากทวีปเอเชียโดยเฉพาะผู้ใช้ชาวจีน ข้อมูลเหล่านี้ชี้ว่าข้อมูลภาษาอื่นๆ นั้นก็มีเป็นจำนวนมากขึ้นในอินเทอร์เน็ต แม้ว่าในปัจจุบันภาษาอังกฤษจะมีมากที่สุด

WORLD INTERNET USAGE AND POPULATION STATISTICS

World Regions	Population (2009 Est.)	Internet Users Dec. 31, 2000	Internet Users Latest Data	Penetration (% Population)	Growth 2000-2009	Users % of Table
Africa	991,002,342	4,514,400	**86,217,900**	8.7 %	1,809.8 %	4.8 %
Asia	3,808,070,503	114,304,000	**764,435,900**	20.1 %	568.8 %	42.4 %
Europe	803,850,858	105,096,093	**425,773,571**	53.0 %	305.1 %	23.6 %
Middle East	202,687,005	3,284,800	**58,309,546**	28.8 %	1,675.1 %	3.2 %
North America	340,831,831	108,096,800	**259,561,000**	76.2 %	140.1 %	14.4 %
Latin America/Caribbean	586,662,468	18,068,919	**186,922,050**	31.9 %	934.5 %	10.4 %
Oceania / Australia	34,700,201	7,620,480	**21,110,490**	60.8 %	177.0 %	1.2 %
WORLD TOTAL	6,767,805,208	360,985,492	**1,802,330,457**	**26.6 %**	399.3 %	100.0 %

ตาราง 13.1 จำนวนผู้ใช้อินเทอร์เน็ตทั่วโลก

http://www.internetworldstats.com/stats.htm

[1] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration by Language	Growth in Internet (2000 - 2009)	Internet Users % of Total	World Population for this Language (2009 Estimate)
English	499,213,462	39.5 %	251.7 %	27.7 %	1,263,830,976
Chinese	407,650,713	29.7 %	1,162.0 %	22.6 %	1,373,859,774
Spanish	139,849,651	34.0 %	669.2 %	7.8 %	411,631,985
Japanese	95,979,000	75.5 %	103.9 %	5.3 %	127,078,679
Portuguese	77,569,900	31.4 %	923.9 %	4.3 %	247,223,493
German	72,337,310	75.0 %	161.1 %	4.0 %	96,389,702
Arabic	60,252,100	17.5 %	2,297.7 %	3.3 %	344,139,242
French	57,017,099	16.9 %	375.2 %	3.2 %	337,046,097
Russian	45,250,000	32.3 %	1,359.7 %	2.5 %	140,041,247
Korean	37,475,800	52.7 %	96.8 %	2.1 %	71,174,317
TOP 10 LANGUAGES	1,492,595,035	33.8 %	379.2 %	82.8 %	4,412,415,512
Rest of the Languages	309,735,422	13.2 %	525.3 %	17.2 %	2,355,389,696
WORLD TOTAL	1,802,330,457	26.6 %	399.3 %	100.0 %	6,767,805,208

ตาราง 13.2 ข้อมูลภาษาของผู้ใช้อินเทอร์เน็ตสูงสุด 10 อันดับแรก

(http://www.internetworldstats.com/stats7.htm)

แต่คำถามพื้นฐานที่สุดคือคำถามว่าข้อมูลเว็บเป็นคลังข้อมูลภาษาหรือไม่ หากยึดตามนิยามโดยทั่วไปของคลังข้อมูลภาษาว่าเป็นข้อมูลที่บอกได้ว่าเป็นตัวแทนของภาษาประเภทใด มีขนาดและโครงสร้างที่ชัดเจน หากเป็นเช่นนี้ เว็บก็ไม่ใช่คลังข้อมูลภาษา แต่ Kilgariff and Grenfenstette (2003) ก็แย้งว่านิยามแบบที่ว่ามาเป็นการตอบคำถามว่าคลังข้อมูลภาษาที่ดีควรเป็นอย่างไร ไม่ใช่การตอบคำถามว่าคลังข้อมูลภาษาคืออะไร สำหรับเขาแล้ว คลังข้อมูลภาษาคือข้อมูลตัวบทที่รวบรวมมาด้วยกัน (collection of texts) ซึ่งเว็บก็เป็นคลังข้อมูลภาษาตามความหมายนี้ และความเป็นตัวแทนของภาษา หากไม่ใช่กรณีภาษาเฉพาะทางแล้ว การจะบอกว่าเรามีข้อมูลที่เป็นตัวแทนของภาษาทั่วไป ก็ค่อนข้างตอบยากว่าต้องนำเอาข้อมูลอะไรมาบ้างจึงจะบอกได้ว่าคลังข้อมูลนั้นเป็นตัวแทนของภาษาทั่วไปแล้ว ดังนั้น ถึงเราจะบอกไม่ได้ว่าข้อมูลเว็บเป็นตัวแทนของภาษาอะไร ก็ไม่ใช่เรื่องแปลก เพราะคลังข้อมูลภาษาอื่นๆ ก็บอกไม่ได้จริงๆ ว่าเป็นตัวแทนภาษาอะไร

ความสนใจในการใช้ประโยชน์จากเว็บเป็นคลังข้อมูลภาษาเริ่มต้นจากพวกนักภาษาศาสตร์คอมพิวเตอร์ซึ่งใช้คลังข้อมูลภาษาในการฝึกสอนระบบประมวลผลภาษาของตน ซึ่งก็เป็นที่ยอมรับกันว่า ยิ่งมีข้อมูลใช้สอนมากเท่าไร ระบบก็จะทำงานได้ดีมากขึ้นด้วย และแหล่งข้อมูลที่สะดวกหาได้ง่ายก็ไม่พ้นข้อมูลจากเว็บต่างๆ ซึ่งนักภาษาศาสตร์คอมพิวเตอร์สามารถเขียนโปรแกรมเพื่อดึงข้อมูลต่างๆ โดยอัตโนมัติได้โดยไม่ยาก

ข้อดีของเว็บคือมีข้อมูลมากมายมหาศาล Fletcher (2004) ยกตัวอย่างการอภิปรายที่เกิดใน CLLT Listserv เมื่อปีค.ศ.2003 ว่าได้มีการค้นหารูปแบบ “not as ADJECTIVE as you think” ว่าเมื่อค้นในคลังข้อมูล MICASE ขนาด 1.7 ล้านคำก็จะพบเพียง 2 ตัวอย่าง เมื่อค้นในคลังข้อมูล BNC ขนาด 100 ล้านคำก็พบเพียง 77 ตัวอย่าง แต่เมื่อลองค้นผ่านเว็บด้วย AltaVista จะได้ว่ามีเว็บ 66,328 หน้าเว็บสำหรับ “as * as you think” หรือเมื่อใช้โปรแกรมอย่าง Web concordance คือ KWiCFinder ก็จะสามารถดึงตัวอย่างมา 500 หน้าเว็บออกมาวิเคราะห์ได้โดยไม่ยาก ตาราง 13.1 แสดงตัวอย่างเปรียบเทียบวลีที่พบใน BNC กับที่พบในเว็บในปีต่างๆ ซึ่งจะเห็นความแตกต่างของจำนวนและการเพิ่มของจำนวนอย่างชัดเจน (Kilgariff and Grenfenstette 2003)

อย่างไรก็ตาม ปัญหาของข้อมูลภาษาที่พบในเว็บมีหลายประการ เช่น การระบุข้อมูลผู้เขียนเพราะเราอาจจะไม่สามารถระบุได้ชัดเจนว่าตัวบทที่เห็นนั้นเขียนโดยใคร เป็นผู้ชายหรือผู้หญิง อายุเท่าไร เป็นคนชาติไหน ซึ่งส่งผลถึงความน่าเชื่อถือของข้อมูลด้วยเพราะข้อความที่เห็นอาจจะไม่ได้เขียนโดยผู้ที่มีความรู้เพียงพอในเรื่องนั้น อาจเป็นข้อมูลที่ผิดมีการใช้ศัพท์ที่ไม่เป็นมาตรฐานยอมรับกันหรือมีการเขียนที่ไม่ถูกไวยากรณ์ ข้อความในรูปภาษาที่เห็นก็อาจจะไม่ได้เขียนโดยเจ้าของภาษาก็ได้ หากคิดว่าจะใช้วิธีการเลือกคัดข้อมูลเฉพาะจากเว็บที่มีที่

อยู่ในพื้นที่ที่ใช้ภาษานั้นเป็นภาษาแม่ ก็ไม่สามารถระบุได้เสมอไปว่าที่อยู่บนเว็บที่เห็นนั้นมาจากสถานที่ตั้งไหน

Sample phrase	BNC (100 M)	WWW fall 1998	WWW fall 2001	WWW spring 2003
medical treatment	414	46,064	627,522	1,539,367
prostate cancer	39	40,772	518,393	1,478,366
deep breath	732	54,550	170,921	868,631
acrylic paint	30	7,208	43,181	151,525
perfect balance	38	9,735	35,494	355,538
electromagnetic radiation	39	17,297	69,286	258,186
powerful force	71	17,391	52,710	249,940
concrete pipe	10	3,360	21,477	43,267
upholstery fabric	6	3,157	8,019	82,633
vital organ	46	7,371	28,829	35,819

Table 1
Frequencies of English phrases in the BNC and on Altavista in 1998 and 2001, and on AlltheWeb in 2003. The counts for the BNC and Altavista are for individual occurrences of the phrase. The counts for AlltheWeb are page counts (the phrase may appear more than once on any page.)

ตาราง 13.3 เปรียบเทียบผลการค้นวลีต่างๆใน BNC และในเว็บ

ความซ้ำซ้อนของข้อมูลก็เป็นอีกปรากฏการณ์หนึ่งบนเว็บที่ตัวบทหนึ่งอาจมีการคัดลอกไว้อยู่หลายที่ จึงอาจจะต้องตรวจสอบข้อมูลที่ดึงจากเว็บมาด้วยว่าไม่เป็นข้อมูลที่ซ้ำกัน นอกจากนี้ ข้อมูลที่ปรากฏบนหน้าเว็บทั้งหลายก็จะมีข้อมูลอื่นๆ ที่ไม่ใช่ข้อความต่อเนื่องที่นักภาษาศาสตร์ต้องการ จึงอาจจะต้องมีการกรองเอาข้อมูลเฉพาะส่วนข้อความภาษาจริงๆ

ปัญหาว่าข้อมูลภาษาเป็นข้อมูลภาษาประเภทใด ก็เป็นประเด็นที่มีผู้สนใจศึกษา Lawrence and Giles (1999) ศึกษาสำรวจสัดส่วนของภาษาแบบต่างๆบนหน้าเว็บ ได้พบว่าส่วนใหญ่เป็นเว็บเชิงพาณิชย์ (83%) การศึกษา (6%) และที่เหลือเป็นประเภทอื่นๆ ส่วน Ide et al. (2002) สนใจตอบคำถามที่มีผู้สงสัยว่า ในเมื่ออินเทอร์เน็ตมีข้อมูลภาษาอังกฤษอยู่มากมายแล้ว ทำไมจึงจำเป็นต้องสร้างคลังข้อมูล American National Corpus อีก โดยได้พยายามคัดเลือกตัวบทจากอินเทอร์เน็ตที่เป็นภาษาอังกฤษแบบอเมริกันและไม่มีข้อจำกัดด้านลิขสิทธิ์ (เลือกจากเว็บไซ้ต์ที่ลงท้ายด้วย .edu หรือ .org) โดยเลือกเว็บที่มีข้อความไม่ต่ำ

กว่า 2,000 คำ หลังจากได้ข้อมูลกว่า 2 ล้านคำจึงนำข้อมูลมาผ่านโปรแกรมกำกับหมวดคำและวิเคราะห์ตัวบทของ Biber et al. (1998) เพื่อจำแนกตัวบทตามมิติต่างๆ 5 มิติที่กำหนดไว้ โดยใช้ลักษณะภาษาที่ปรากฏในตัวบทนั้น ผลการวิเคราะห์พบว่าภาษาที่ได้จากเว็บไม่ได้มีความหลากหลายแบบที่ต้องการในการสร้าง ANC ตัวบทที่ได้ส่วนใหญ่เป็นงานเขียนที่เป็นทางการ ทางวิชาการและเป็นบทสำหรับกล่าวสุนทรพจน์ จึงไม่มีตัวบทอื่นๆ ที่ต้องการ เว็บจึงไม่มีความหลากหลายเท่าภาษาที่เกิดขึ้นภายนอก

3. การรวบรวมข้อมูลเว็บสร้างเป็นคลังข้อมูลภาษา

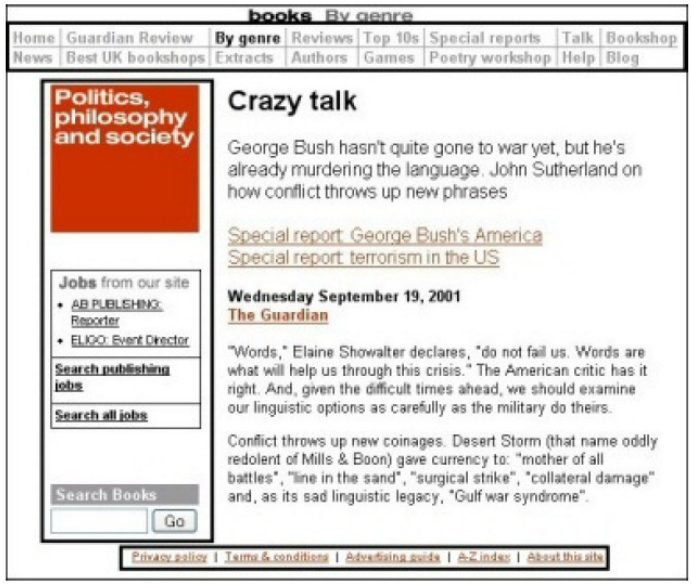
แม้ว่าข้อมูลภาษาที่อยู่บนหน้าเว็บต่างๆ จะมีข้อจำกัดหลายประการตามที่ได้กล่าวมา แต่ในกรณีที่เราไม่มีคลังข้อมูลภาษาที่ต้องการ อินเทอร์เน็ตก็เป็นแหล่งที่เราสามารถหาข้อมูลมาเป็นตัวอย่างของการศึกษาภาษาได้ Fletcher (2004) ได้รายงานการทดลองสร้างคลังข้อมูลภาษาด้วยวิธีการรวบรวมตัวบทจากหน้าเว็บต่างๆ ขึ้นมา เขาได้สร้างเครื่องมือที่จำเป็นเพื่องานนี้

ในการคัดเลือกเว็บนั้น ถ้าเรากำหนดคำค้นใดเราก็จะได้ข้อมูลที่มีคำค้นนั้นซึ่งก็จะทำให้ข้อมูลที่รวบรวมมาถูกจำกัดด้วยคำที่ใช้ วิธีที่จะเลี่ยงเพื่อให้ได้ข้อมูลที่ไม่เจาะจงอาจทำได้โดยการใช้คำค้นที่เป็นคำไวยากรณ์ เช่น the หรือ of หรือ a เป็นต้น แต่ในการตัดสินใจว่าจะดึงข้อมูลจากเว็บไหนมารวบรวมไว้นั้นยังต้องพิจารณาปัจจัยอื่นๆ ได้แก่ พิจารณาความยาวของหน้าเว็บซึ่งก่อนที่จะดาวน์โหลดข้อมูลมาเราสามารถพิจารณากำหนดขนาดของเว็บที่เหมาะสม ไม่เอาข้อมูลจากหน้าเว็บที่มีขนาดเล็กเกินไปเพราะมักไม่มีข้อความต่อเนื่องที่จะเป็นประโยชน์แก่การศึกษาปัญหาและมักจะมีสัดส่วนข้อมูลที่ไม่ต้องการสูง ในขณะเดียวกัน เว็บที่มีขนาดใหญ่มากๆ ก็อาจเป็นหน้าที่รวมรายการต่างๆ รายการ

ศัพท์ซึ่งก็ไม่ใช่สิ่งที่เราอยากจะได้ จึงควรกำหนดขอบเขตขนาดที่ควรจะดึงข้อมูลหน้าเว็บได้

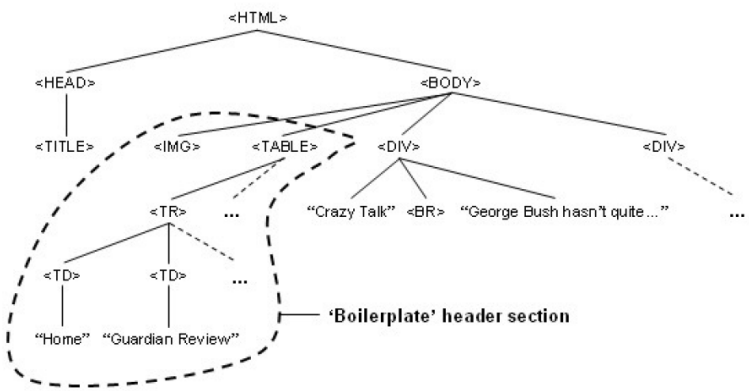
เมื่อพิจารณาหาวิธีกรองข้อมูลส่วนที่ไม่ต้องการ คือส่วนที่ไม่ใช่ข้อความต่อเนื่องแต่เป็นโฆษณาหรือเมนูต่างๆบนหน้าเว็บออกไปด้วย Fletcher (2004) เรียกข้อมูลส่วนนี้ว่าเป็น noise ซึ่ง Fletcher มองว่าข้อมูล noise เหล่านี้มีลักษณะซ้ำๆ และสามารถระบุรูปแบบได้เป็นเหมือน boilerplate นอกจากนี้ ข้อมูลบางส่วนบนเว็บก็เป็นข้อมูลที่ซ้ำกันอาจจะเนื่องมาจากมีการคัดลอกไปไว้อีกเว็บ หรือเป็นการทำซ้ำ (mirror) ไปไว้อีกที่ วิธีการที่ Fletcher ใช้คือเทคนิค duplicate identical document ซึ่งมีอัลกอลิธึ่มที่จะแปลง document เป็นรหัสของตัวอักษร 16 ตัวที่เป็นรหัสเฉพาะของ document นั้น ในการโหลดข้อมูลเข้ามาจึงมีการเทียบรหัสที่วิเคราะห์ได้นี้ว่าซ้ำกับของเดิมหรือไม่ แต่ก็มีข้อมูลประเภทที่ไม่ใช่เป็นการทำซ้ำของเดิม แต่เป็นการปรับข้อมูลให้มีความต่างซึ่งวิธีการที่กล่าวมาจะระบุไม่ได้ จึงใช้วิธีเทียบความคล้ายโดยดูที่ขนาดว่าแตกต่างกันไม่เกิน 5% ดูรายการความถี่ของ 3-gram ที่ได้จากในเอกสารทั้งสอง ถ้าใน 25 รายการที่มีความถี่สูงสุดมี 20 รายการที่มีรูปคำและความถี่คล้ายกัน ก็ถือว่าเป็นข้อมูลที่ซ้ำซ้อนกัน

ความซ้ำอาจเป็นเรื่องภายในเอกสารนั้นเองก็ได้ วิธีหนึ่งที่ใช้ตรวจสอบคือ หา n-gram ที่ยาวมากๆ ในเอกสารนั้น หากพบว่ามีการซ้ำกันของ n-gram นั้นเกิดขึ้นจำนวนหนึ่ง ก็แสดงว่ามีการซ้ำกันของข้อความในเอกสารนั้น ซึ่งอาจเป็นไฟล์ที่เป็น log ของเครื่องแม่ข่าย หรือเป็นการตอบข้อความ (message) โดยมีการคัดลอกข้อความจากข้อความเดิมดังที่เห็นในเว็บบอร์ด



รูป 13.1 แสดงบริเวณ boilerplate ที่พบในหน้าเว็บ(อยู่ในกรอบดำ)

(Kehoe & Gee 2007)



รูป 13.2 แสดงโครงสร้างแท็กในเว็บและส่วนที่เป็น boilerplate

(Kehoe & Gee 2007)

ปัญหาลิขสิทธิ์ก็เป็นอีกปัญหาหนึ่งที่อาจจะต้องคำนึงถึง วิธีแก้คือเสี่ยงการนำตัวบทบนหน้าเว็บที่มีการสงวนลิขสิทธิ์ไว้ เช่น หากมีคำว่า copyright หรือ all rights reserved ปรากฏอยู่ ก็จะเลี่ยงไม่นำข้อมูลจากเว็บนั้นมารวบรวมไว้

การทดลองของ Fletcher ที่ใช้วิธีการต่างๆกรองข้อมูลที่ต้องการมา ทำให้เขาสามารถสรุปขนาดของไฟล์ได้ว่า ไฟล์ขนาดที่ใหญ่กว่า 200 KB นั้นใช้ไม่ได้เลย ที่มีขนาด 100 KB ก็ใช้ได้เพียง 10 ไฟล์ ขนาดที่ใช้ได้ดีสุดคือประมาณ 50 KB ผลจากงานนี้จึงนำไปใช้กับโปรแกรม KWiCFinder ให้คัดเลือกโหลดข้อมูลที่มีขนาดในช่วง 10-150 KB ซึ่งเมื่อเทียบกับงานของ Cavaglia and Kilgarriff (2001) ซึ่งใช้วิธีการทางสถิติเปรียบเทียบรายการความถี่คำในหน้าเว็บกับที่พบใน BNC หรือของ Ide et al (2002) ในการสร้าง ANC ซึ่งกำหนดตัวบทให้มีขนาดอย่างต่ำ 2,000 คำ หากยึดตามแนวปฏิบัตินี้ เว็บจำนวนมากก็จะไม่ถูกนำมาใช้ เพราะหน้าเว็บจำนวนมากเลยที่จะทำเป็นข้อความที่ไม่ยาวเกินไป

ข้อดีของการรวบรวมข้อมูลจากเว็บมาสร้างเป็นคลังข้อมูลคือทำให้เรามีข้อมูลที่สามารถนำมาตรวจสอบภายหลัง สามารถทำซ้ำการศึกษาใดๆ ได้เพราะยังคงมีข้อมูลชุดเดิมให้ทดสอบอยู่ เพราะหากใช้ข้อมูลจากเว็บโดยตรง การค้นแต่ละครั้งในวันที่ต่างกันก็อาจให้ผลที่แตกต่างกันได้เพราะข้อมูลมีการปรับเปลี่ยนอยู่เสมอ แต่บ้างก็มองตรงข้าม คือมองว่าการที่ทำการศึกษาซ้ำจากข้อมูลบนเว็บโดยตรงแล้วได้ผลลักษณะเดิม แสดงว่าผลการศึกษานั้นน่าเชื่อถือ เป็นการยืนยันที่ดี เพราะได้ทดสอบกับข้อมูลที่มาจากเว็บในเวลาที่ต่างกันเหมือนมาจากต่างส่วนของอินเทอร์เน็ตแล้ว

4. การใช้เว็บเป็นคลังข้อมูลภาษา

ในกรณีที่จำเป็น เราก็สามารถใช้ประโยชน์จากอินเทอร์เน็ตเพื่อการศึกษาภาษาได้โดยไม่จำเป็นต้องรวบรวมหรือคัดเลือกข้อมูลจากหน้าเว็บมาใช้ โดยเฉพาะในกรณีค้นหาตัวอย่างการใช้ซึ่งคลังข้อมูลภาษาที่มีอยู่ไม่สามารถให้จำนวนตัวอย่างได้เพียงพอ เพราะข้อมูลในเว็บมีจำนวนมหาศาลจึงเป็นที่ที่จะช่วยหาตัวอย่างที่พบน้อยๆ เช่นนั้นได้ดี หรือในกรณีของคำที่สร้างขึ้นใหม่ เช่น ตุลาการภิวัตน์ ประชาภิวัตน์ คลังข้อมูลที่สร้างขึ้นจากข้อมูลภาษาก่อนปีพ.ศ. 2548 ก็จะไม่มีข้อมูลคำเหล่านี้

อินเทอร์เน็ตยังเป็นแหล่งสำหรับตรวจสอบคำสะกดได้ด้วย ในกรณีที่เราเกิดความไม่แน่ใจ Kilgarriff and Grefenstette (2003) ยกตัวอย่างคำว่าไม่แน่ใจว่าจะสะกด speculator หรือ speculater ซึ่งเมื่อค้นจากเว็บก็จะพบว่าคำแรกนั้นมีพบใช้มากกว่าคือพบ 82,000 รายการ .ในขณะที่คำหลังพบเพียง 67 รายการ จึงสรุปได้ว่าเป็นการสะกดแบบแรก หรือในตัวอย่างภาษาไทย เราอาจไม่แน่ใจว่าจะเขียน *โลกาภิวัตน์* หรือ *โลกาภิวัฒน์* เมื่อทดลองค้นด้วย Google ในวันที่ 9 กันยายน พ.ศ.2551 จะพบว่า *โลกาภิวัตน์* มีรายการพบประมาณ 290,000 รายการ ส่วน *โลกาภิวัฒน์* พบจำนวนประมาณ 264,000 รายการ แสดงว่าน่าจะเป็นคำแรกซึ่งก็เป็นคำที่ตรงกับที่ระบุไว้ในพจนานุกรมราชบัณฑิต แต่จำนวนการใช้ที่ใกล้เคียงกันนี้แสดงให้เห็นว่าเป็นคำที่ใช้สับสนและมีผู้ใช้ทั้งสองแบบพอๆกัน และเมื่อทดลองค้นทั้งสองคำนี้ในวันที่ 12 กันยายน พ.ศ.2552 กลับพบว่า *โลกาภิวัตน์* มีจำนวนหน้าเว็บ 240,000 รายการขณะที่ *โลกาภิวัฒน์* พบ 2,510,000 รายการ ซึ่งแสดงให้เห็นการเปลี่ยนแปลงที่เกิดขึ้นของคำนี้อย่างชัดเจน

อีกตัวอย่างคือ คำว่า *กฎมนเทียรบาล* ในพจนานุกรมฉบับราชบัณฑิตยสถาน เมื่อค้นด้วย Google ในวันที่ 9 กันยายน พ.ศ.2551 จะพบ

1,810 รายการ ซึ่งมากกว่าคำที่สะกดผิดอย่าง *กฎมลเทียรบาล*ที่พบเพียง 341 รายการ และก็มากกว่าคำที่พจนานุกรมว่าเป็นคำโบราณ *กฎมณเทียรบาล* (พบ 433 รายการ) หรือ *กฎมนเฑียรบาล* (พบ 793 รายการ) อย่างไรก็ตาม จากการค้นด้วย Google ในวันที่ 12 กันยายน พ.ศ.2552 กลับพบว่า คำที่มีอยู่ในพจนานุกรม *กฎมนเทียรบาล* พบเพียง 3,990 รายการ น้อยกว่าคำที่ไม่มีในพจนานุกรม คือ *กฎมลเฑียรบาล* (11,100 รายการ) และน้อยกว่าคำที่พจนานุกรมว่าเป็นคำโบราณคือ *กฎมณเทียรบาล* (19,900 รายการ) และ *กฎมนเฑียรบาล* 6,940 รายการ) ซึ่งแสดงให้เห็นแนวโน้มว่า การสะกดคำเหล่านี้กำลังมีการเปลี่ยนแปลงไม่เป็นไปตามมาตรฐานที่กำหนด

ผลที่ได้จากการค้นนั้นเป็นการประมาณปริมาณการใช้ที่พบ ซึ่งก็เป็นไปได้ว่าคำที่พบใช้มากกว่าอาจไม่เป็นไปตามที่บัญญัติไว้ก็ได้ เช่น คำว่า *มุกตลก* ซึ่งสะกดตามพจนานุกรมราชบัณฑิตยสถานเมื่อค้นด้วย Google ในวันที่ 9 กันยายน พ.ศ.2551 จะพบประมาณ 46,600 รายการ ส่วน *มุขตลก* จะพบมากกว่าคือพบ 211,000 รายการ และเมื่อค้นด้วย Google ในวันที่ 12 กันยายน พ.ศ.2552 จะได้จำนวน 179,000 และ 286,000 ตามลำดับ ซึ่งก็ยังแสดงว่า *มุกตลก* มีการใช้ที่น้อยกว่าแต่ก็จะเห็นมีอัตราการใช้ที่เพิ่มที่สูงขึ้นอย่างมาก กรณีนี้แสดงให้เห็นถึงปัญหาของการใช้ของคนส่วนใหญ่ว่าไม่ตรงตามที่บัญญัติ พจนานุกรมฉบับมติชนจึงเลือกบรรจุคำ *มุขตลก* ไว้ แต่ข้อโต้แย้งเรื่องการใช้คำสองคำนี้ที่ปรากฏเป็นข่าว ก็อาจจะทำให้จำนวนการใช้ของ *มุกตลก* สูงมากขึ้นในหนึ่งปีถัดมา

คำว่า *โชห่วย* ซึ่งควรเป็นคำที่พบมาก แต่เมื่อค้นด้วย Google (ในวันที่ 9 กันยายน พ.ศ.2551) ก็พบประมาณ 79,600 รายการ มากกว่าคำ *โชวห่วย* เล็กน้อย (พบ 69,300 รายการ) แต่น้อยกว่าคำ *โชว์ห่วย* ซึ่งพบถึง 283,000 รายการ กรณีเช่นนี้ การตีความใดๆ ก็จะต้องระวัง เพราะว่าคำหลังนั้นมีการใช้ใน

ความหมายของการแสดงที่แย่ด้วย ทั้งสามคำนี้เมื่อนำมาค้นด้วย Google อีกครั้งในวันที่ 12 กันยายน พ.ศ.2552 ก็พบจำนวนที่แตกต่างออกไป *โชห่วย* (2,000,000 รายการ) *โชวห่วย* (1,500,000 รายการ) และ *โชว์ห่วย* (1,760,000 รายการ) กรณีตัวอย่างต่างๆ ที่ยกมานี้เป็นตัวอย่างที่ทำให้เราต้องระวังในการตีความข้อมูลที่พบ จำนวนการใช้ที่แตกต่างกันมากอาจช่วยให้เราสรุปได้ว่าควรจะเลือกสะกดคำแบบใด แต่การสะกดที่พบอาจเป็นการสะกดตามความนิยมซึ่งไม่ตรงตามที่บัญญัติไว้ก็ได้ และจำนวนการใช้ที่ไม่แตกต่างกันชัดเจนก็เป็นการบ่งชี้ว่าคำที่วิเคราะห์อยู่นั้นมีความสับสนในการใช้อยู่พอสมควร การเก็บข้อมูลจำนวนการใช้ในช่วงเวลาที่แตกต่างกัน สามารถช่วยให้เห็นแนวโน้มการเปลี่ยนแปลงของการใช้คำได้ในระดับหนึ่ง

นอกจากนี้ กรณีที่คำมีหลายความหมาย (polysemous) การค้นคำจะต้องอาศัยวิธีการค้นแบบซับซ้อนขึ้น เช่น คำว่า chair ถ้าเราไม่ต้องการข้อมูลที่มีคำ chair ที่หมายถึงเก้าอี้ ก็อาจจะต้องกำหนดคำที่ช่วยกันข้อมูลที่ไม่ต้องการ เช่นระบุว่าจะไม่เอาข้อมูลที่มีคำ furniture หรือ table ด้วย หรือไม่ก็เพิ่มคำค้นเพิ่มเติมว่าจะต้องมีคำว่า committee ด้วยเพื่อให้ได้ chair ที่หมายถึงประธาน แต่ก็ยังมีข้อจำกัดอีกอย่างหนึ่งที่พบในการค้นข้อมูลภาษาจากเว็บ คือ ข้อมูลเหล่านี้ไม่มีการกำกับข้อมูลภาษาอย่างเช่นหมวดคำเอาไว้ สิ่งที่ค้นได้จึงเป็นการค้นจากรูปคำเพียงอย่างเดียว คำว่า chair ที่ค้นได้เราไม่อาจบอกได้ว่าเป็นคำที่เป็นคำนามหรือคำกริยา การจะหาข้อสรุปก็จำเป็นต้องดึงตัวอย่างที่ค้นมาได้มาวิเคราะห์จำนวนหนึ่งเพื่อหาค่าสถิติสำหรับประมาณการต่อไป

การค้นหาโครงสร้างวากยสัมพันธ์ แม้ว่าข้อมูลบนเว็บจะไม่มีการกำกับหมวดคำไว้ แต่การค้นโครงสร้างวากยสัมพันธ์บางอย่างก็สามารถทำได้หากโครงสร้างนั้นสามารถกำหนดโดยรูปคำได้ เช่น เมื่อต้องการค้นคำ relative

pronoun *who* ในตำแหน่งที่ควรเป็น *whom* ถ้าค้น *who* กับ *whom* โดยตรงก็จะได้ผลออกมามากและเราก็ไม่รู้ว่า *who* ที่พบใช้เป็นกรรมหรือไม่ เราจึงควรค้นโดยกำหนดรูปแบบการค้นให้ชัดขึ้นอย่าง *who I like, whom I like* ด้วยวิธีนี้เราก็สามารถนำผลที่ได้มาเทียบได้ว่ามีการใช้ *who* แทน *whom* ในสัดส่วนเท่าใด และเวลาค้นจริงๆ เราก็สามารถสร้างรูปแบบหลายๆแบบ เช่น *who I got to know, whom I got to know; who I called, whom I called; who I gave, whom I gave* เป็นต้น ผลที่ได้จะทำให้เราสามารถศึกษาเปรียบเทียบการใช้ *who* กับ *whom* ได้ว่า ในบริบทเช่นใดที่ *who* ถูกใช้แทน *whom* ได้ และบริบทแบบใดที่ *whom* ยังคงเป็นที่นิยมใช้ เป็นเรื่องของกริยาที่เกี่ยวข้องหรือไม่ เป็นต้น (ดู Meyer et al. 2003)

ส่วนการศึกษาในเชิงการเปลี่ยนแปลง ข้อมูลจากเว็บอาจใช้ไม่ได้โดยตรง เหตุผลแรกเป็นเรื่องของเว็บเองที่เพิ่งมีใช้ในช่วงไม่เกิน 20 ปี และข้อมูลที่ค้นได้ก็ไม่ได้บอกเวลาของการสร้างตัวบทโดยตรง แต่เราก็อาจศึกษาในเชิงการเปลี่ยนแปลงได้โดยค้นในช่วงเวลาที่ต่างกันจริง เช่น เทียบข้อมูลที่เปลี่ยนไปในแต่ละปี ดังตัวอย่างที่กล่าวมาแล้วข้างต้น แต่นั่นก็หมายความว่าเราต้องทำการศึกษาแบบต่อเนื่องระยะยาวด้วย

5. Web-based **คอนคอร์แด๊นซ์**

โปรแกรม web-based คอนคอร์แด๊นซ์ คือโปรแกรมคอนคอร์แด๊นซ์ประเภทหนึ่งที่ใช้ข้อมูลจากเว็บต่างๆ มาเป็นข้อมูลพื้นฐานในการสืบค้น แตกต่างจากโปรแกรมตัวสืบค้น (search engine) ที่เมื่อโปรแกรมค้นเว็บที่มีข้อมูลตรงตามที่ต้องการได้แล้วจะแสดงผลในรูปแบบที่เป็น kwic ไม่ใช่แสดงเป็นรายการ

หน้าเว็บแบบโปรแกรมตัวสืบค้น ได้แก่ WebCorp[2] ของ University of Liverpool ซึ่งเป็นบริการผ่านเว็บให้เลือกใช้ตัวสืบค้น (search engine) อย่าง AltaVista, Google, Metaclawer แล้วแสดงผลการสืบค้นในรูปแบบของ kwic และโปรแกรม KWiCFinder[3] ของ William Fletcher ซึ่งเป็นโปรแกรมติดตั้งบนคอมพิวเตอร์ให้สามารถเรียกใช้ตัวสืบค้นจาก AltaVista และนำผลที่สืบค้นได้มาแสดงในรูปที่เป็น kwic อย่างไรก็ตาม โปรแกรมเหล่านี้ต่างก็ต้องอาศัยโปรแกรมตัวสืบค้นบนอินเทอร์เน็ตเป็นพื้นฐานคือจะเข้าถึงเว็บที่โปรแกรมตัวสืบค้นได้มีการทำดัชนีไว้แล้ว Meyer et al. (2003) กล่าวถึงการเปรียบเทียบการทำงานของตัวสืบค้นต่างๆ ได้แก่ Google, Fast, AltaVista, และ N.Light ซึ่งผลที่ได้จากการสืบค้นคำเดียวกันจากแต่ละโปรแกรมนั้นแม้จะมีจำนวนไม่เท่ากัน แต่ถ้ามองเทียบเป็นสัดส่วนแต่ละโปรแกรมจะให้ผลใกล้เคียงกัน เช่น เมื่อทดลองค้นคำ *chairman, chairperson,* และ *chairwoman* ในช่วงเวลาต่างกันดังรูป *chairman* จะเป็นคำที่พบมากสุดจากทุกตัวสืบค้น (ตัวเลขที่ได้เป็นจำนวน hit หรือหน้าเว็บที่รายงานออกมา แต่ไม่ใช่จำนวนคำนั้นโดยตรง)

---

[2] http://www.webcorp.org.uk/

[3] http://www.kwicfinder.com/

	Google	Fast	AltaVista	N. Light
29-3-2002				
chairman	7,710,000 (85%)	4,325,141 (86%)	3,506,210 (87%)	2,063,336 (95%)
chairperson	1,190,000 (13%)	644,491 (13%)	444,227 (11%)	30,017 (1%)
chairwoman	141,000 (2%)	63,735 (1%)	65,312 (2%)	80,495 (4%)
Total	9,041,000	5,033,367	4,015,749	2,173,848
4-4-2002				
chairman	7,710,000 (85%)	4,325,166 (86%)	3,020,596 (87%)	2,081,540 (95%)
chairperson	1,190,000 (13%)	644,491 (13%)	391,205 (11%)	30,255 (1%)
chairwoman	141,000 (2%)	63,876 (1%)	59,667 (2%)	80,733 (4%)
Total	9,041,000	5,033,533	3,471,468	2,192,528
15-5-2002				
chairman	7,810,000 (86%)	6,585,252 (86%)	3,023,125 (87%)	2,156,782 (95%)
chairperson	1,170,000 (13%)	955,564 (12%)	391,262 (11%)	31,176 (1%)
chairwoman	147,000 (2%)	116,245 (2%)	59,786 (2%)	82,789 (4%)
Total	9,127,000	7,657,061	3,474,173	2,270,747

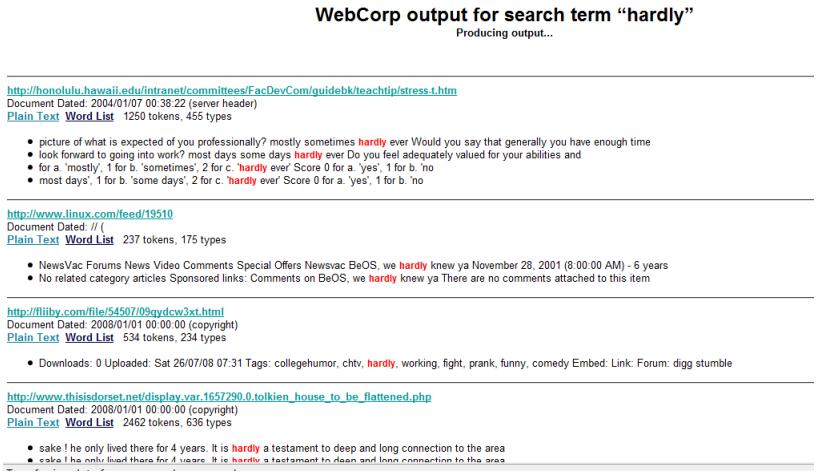
ตาราง 13.4 เปรียบเทียบผลการสืบค้นจากตัวสืบค้นต่างๆ

(Meyer et al. 2003)

5.1 โปรแกรม WebCorp

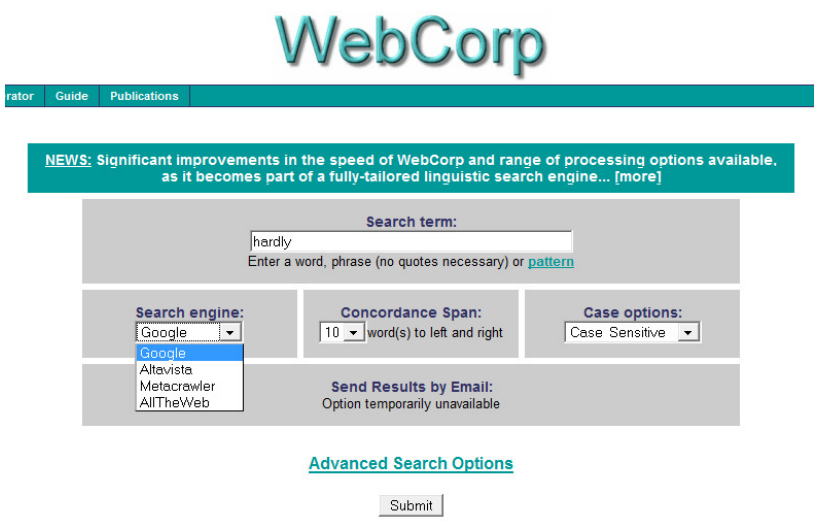
WebCorp เป็นโครงการที่พัฒนาที่ University of Liverpool ตั้งแต่เดือนธันวาคม ปีค.ศ.2000 ขั้นตอนการทำงานของ WebCorp มี 6 ขั้นตอน เริ่มจากการรับคำสั่งค้นจากผู้ใช้ จากนั้นจะแปลงและส่งคำสั่งนั้นให้ตัวสืบค้น (search engine) ตัวสืบค้นจะหาเว็บที่ตรงกับคำสั่งค้นและส่งรายการ URL กลับมา โปรแกรม WebCorp จะประมวลผลตัวบทที่ได้และแสดงผลออกมาในรูปแบบของคอนคอร์แด๊นซ์ให้กับผู้ใช้ดังตัวอย่างในรูป คือ แจงเป็นรายการหน้าเว็บต่างๆ

พร้อมทั้งแสดงข้อความที่มีคำค้นปรากฏอยู่ตรงกลาง (หรือจะสั่งให้แสดงในรูปที่เป็น kwic ก็ได้) ผู้ใช้ยังสามารถกดดูหน้าเว็บนั้นผ่านทางลิ้งค์ที่ให้



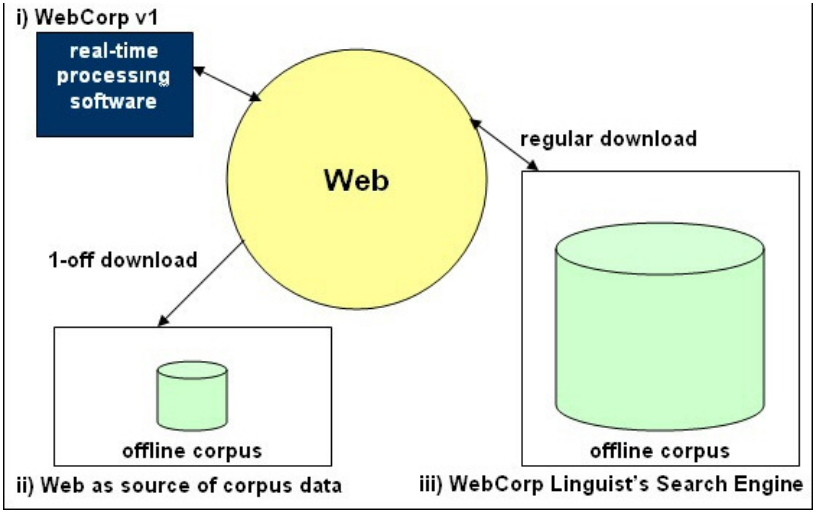
รูป 13.3 ผลลัพธ์จากการค้นคำ hardly ผ่านทาง WebCorp

ผู้ใช้สามารถเลือกได้ว่าจะใช้ตัวสืบค้นใด ได้แก่ Google, AltaVista, metaclawler บริบทซ้ายขวาที่ต้องการยาวเพียงใด ดังแสดงในรูป แต่ปัญหาหนึ่งของโปรแกรมนี้คือเวลาในการทำงาน ซึ่งการค้นหาคำที่ต้องการจนได้ผลแสดงออกมานั้นใช้เวลาค่อนข้างนานหลายนาที แต่ข้อดีก็คือมีการเพิ่มวิธีการค้นให้เหมือนโปรแกรมคอนคอร์แด็นซ์เช่นมีการใช้ wildcard character อย่างเช่น * ได้ ซึ่งโปรแกรมตัวสืบค้นโดยปกติทำไม่ได้ WebCorp จึงเป็นความพยายามที่จะต่อยอดโปรแกรมตัวสืบค้นทั้งหลายให้ทำงานค้นข้อมูลในแบบที่นักภาษาศาสตร์ต้องการ



รูป 13.4 หน้าจอการสืบค้นจาก WebCorp

ในปัจจุบัน WebCorp พยายามปรับแนวการใช้เว็บเป็นคลังข้อมูล จากเดิมที่ทำเป็นโปรแกรมไปค้นข้อมูลจากเว็บโดยตรงคือแนวทางแรกที่ใช้เว็บเป็นคลังข้อมูลโดยตรง ก็เปลี่ยนไปสู่แนวทางที่สองที่มีการโหลดข้อมูลที่ต้องการจากเว็บมาเก็บไว้เองซึ่งเป็นแนวการสร้างคลังข้อมูลขึ้นจากข้อมูลเว็บ แต่แบบนี้ก็ทำให้ข้อมูลมีลักษณะที่หยุดนิ่ง จึงมีอีกแนวทางหนึ่งคือโหลดข้อมูลที่ต้องการเข้ามาอยู่ตลอดเวลาเพื่อให้ได้ข้อมูลที่ปรับเพิ่มตลอดเวลา ดังแสดงเปรียบเทียบทั้งสามแนวทางในรูปข้างล่าง

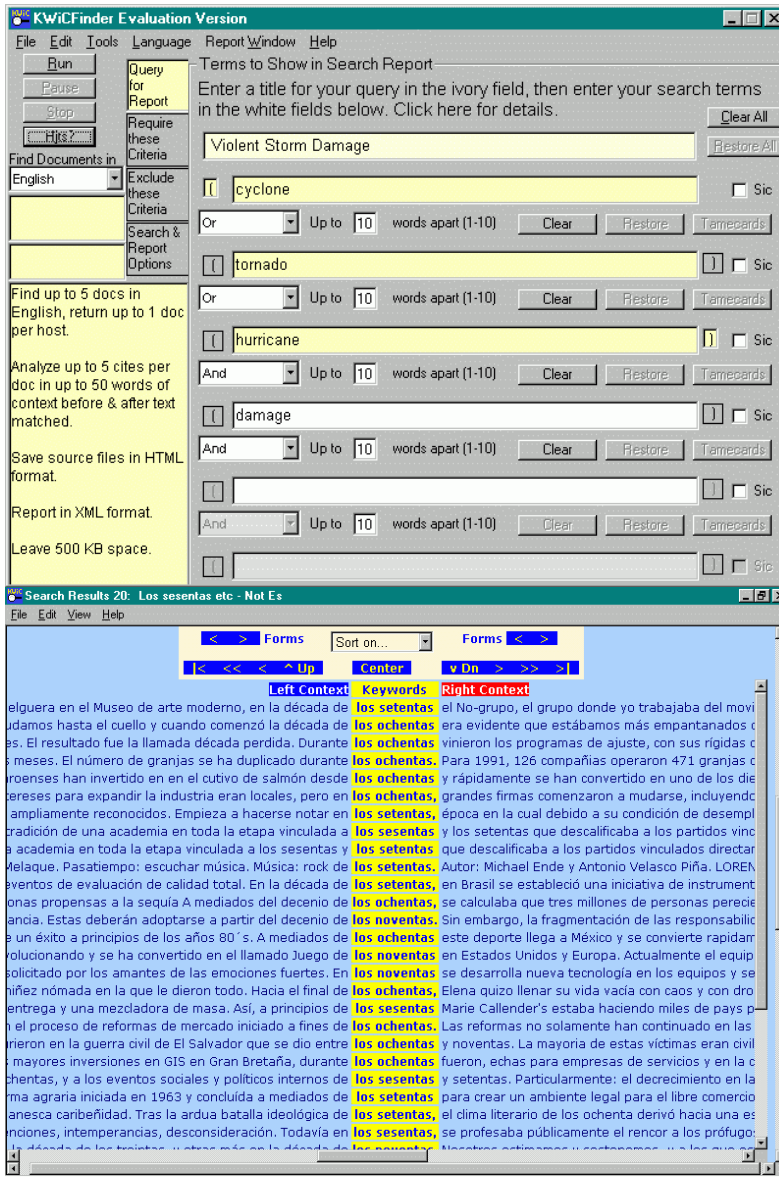


รูป 13.5 แสดงแนวทางการปรับ WebCorp

(Kehoe and Gee, 2007)

5.2 โปรแกรม KWiCFinder

เป็นโปรแกรมที่พัฒนาโดย William Fletcher ซึ่งเป็นโปรแกรมสำหรับติดตั้งบนคอมพิวเตอร์ เมื่อเรียกโปรแกรมมาใช้งานโปรแกรมจะสามารถเรียกใช้ตัวสืบค้นจาก AltaVista และนำผลที่สืบค้นได้มาแสดงในรูปต่างๆ รูปข้างล่างแสดงตัวอย่างหน้าจอของโปรแกรมซึ่งผู้ใช้สามารถกำหนดตัวเลือกการค้นย่อยลงไปอีกได้ (สามารถดูหน้าจอตัวเลือกต่างๆจากเว็บไซ้ต์ของโปรแกรมนี้เอง) เมื่อค้นได้ผลที่ต้องการแล้ว โปรแกรมสามารถแสดงผลได้ทั้งในรูปที่เป็น XML format หรือจะแสดงเป็นรูป kwic แบบโปรแกรมคอนคอร์แด๊นซ์ทั่วไปก็ได้ หรือจะแสดงเป็นตาราง HTML ก็ได้



รูป 13.6 โปรแกรม KWiCFinder และผลการค้นผ่านทางเว็บ

6. บทสรุป

แม้ว่าจะยังมีข้อถกเถียงเรื่องข้อมูลภาษาในอินเทอร์เน็ตว่าสามารถมองและใช้เป็นคลังข้อมูลภาษาแบบที่ใช้กันได้หรือไม่ แต่เว็บก็เป็นแหล่งข้อมูลภาษาขนาดใหญ่ที่สามารถนำมาใช้ประโยชน์ในเรื่องต่างๆ ได้ เช่น สำรวจการเขียนคำวลี แบบต่างๆ ค้นหาตัวอย่างคำยากๆ ที่คลังข้อมูลภาษาโดยทั่วไปไม่สามารถให้ตัวอย่างได้มากพอ และสามารถใช้บ่งชี้การเปลี่ยนแปลงของภาษาได้หากมีการเปรียบเทียบผลการใช้ในช่วงเวลาต่างๆ แต่ผู้ใช้ก็ต้องเข้าใจข้อจำกัดและความไม่แน่ชัดของข้อมูล ไม่เหมือนกับคลังข้อมูลภาษาทั่วไปที่เรารู้ถึงโครงสร้างและสัดส่วนของตัวบทประเภทต่างๆ ในคลังข้อมูลที่ใช้

บรรณานุกรม

Abe, M. 2003. A Corpus-based Contrastive Analysis of Spoken and Written Learner Corpora: The Case of Japanese-speaking Learners of English. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics*, pp. 1-9. Lancaster University.

Abney, S. 1996. Statistical Methods And Linguistics. In J. Klavans and P. Resnik (eds.). *The Balancing Act*, Cambridge: MIT Press.

Aijmer, K. 1986. Why is Actually So Frequent In Spoken English. In G. Tottie and I. Bäcklund (eds.). *English in Speech and Writing: A Symposium*, pp. 119-129. Stockholm: Almqvist & Wiksell.

Aijmer, K. 1996. *Conversational routines in English : convention and creativity*. London ; New York: Longman.

Aijmer, K., Altenberg, B. and Svartvik, J. (eds.) 1991. *English corpus linguistics : studies in honour of Jan Svartvik*. London ; New York: Longman.

Akkerman, E. 1984. Verb and Particle Combinations: Particle Frequency Ratings and Idiomaticity. *ICAME News* 8: 60-70.

Allen, J. 1995. *Natural language understanding*. Redwood City, Calif.: Benjamin/Cummings Pub. Co.

Altenberg, B. 1993. Recurrent Verb-Complement Constructions in the London-Lund Corpus. In J. Aarts, P. d. Haan and N. Oostdijk (eds.). *English Language Corpora: Design Analysis and Exploitation*, pp. 227-245. Amsterdam: Rodopi.

Altenberg, B. 1994. On the Functions of such in Spoken and Written English. In N. Oostdijk and P. d. Haan (eds.). *Corpus-based Research into Language*, pp. 223-240. Amsterdam: Rodopi.

Altenberg, B. and Eeg-Olofsson, M. 1990. Phraseology in Spoken English: Presentation of a

Project. In J. Aarts and W. Meijs (eds.). *Theory and Practice in Corpus Linguistics*, Amsterdam and Atlanta: Rodopi.
Altenberg, B. and Granger, S. 2001. The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics* 22(2): 173-194.
Arrieta, B., Ilarraza, A. D. d., Gojenola, K., Maritxalar, M. and Oronoz, M. 2003. A database system for storing second language learner corpora. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics*, pp. 33-41. Lancaster University.
Aston, G. 1995. Corpora in Language Pedagogy: Matching Theory and Practice. In G. Cook and B. Seidlhofer (eds.). *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press.
Aston, G., Bernardini, S., Stewart, D. and Bernadini, S. 2004. *Corpora And Language Learners*. John Benjamins Publishing Co.
Atkinson, D. 1992. The Evolution of Medical Research Writing From 1735 to 1985: The case of the Edinburgh Medical Journal. *Applied Linguistics* 13: 337-374.
Baker, M. 1992. *In other words : a coursebook on translation*. London ; New York: Routledge.
Baker, M. 1993. Corpus Linguistics And Translation Studies: Implications And Applications. In M. Baker, G. Francis and E. T. Bonelli (eds.). *Text and Technology in honor of John Sinclair*, pp. 233-252. John Benjamins.
Barlow, M. 1995. ParaConc: A Concordancer. *Texts* 10.
Barlow, M. 1996. Corpora for Theory and Practice. *International Journal of Corpus Linguistics* 1(1).
Béjoint, H. 2000. *Modern lexicography: an introduction*. Oxford: Oxford University Press.

Belz, J. A. 2004. Learner corpus analysis and the development of foreign language proficiency. *system* 32(4): 577-591.

Belz, J. A. and Vyatkina, N. 2005. Learner Corpus Research and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles. *Canadian Modern Language Review* 62(1): 17-48.

Bernardini, S. 1997. A 'trainee' translator's perspective on corpora. *Corpus use and learning to translate Bertinoro*.

Berry-Rogghe, G. L. M. 1973. The Computation of Collocations and Their Relevance in Lexical Studies. In A. J. Aitken, R. W. Bailey and N. Hamilton-Smith. (eds.). *The Computer and Literary Studies*, Edinburg: Edinburg University Press.

Berry-Rogghe, G. L. M. 1974. Automatic Identification of Phrasal Verbs. In J. L. Mitchell (ed.). *Computers in the Humanities*, Edinburgh: Edinburgh University Press.

Biber, D. 1988. *Variation across speech and writing*. Cambridge [Cambridgeshire] ; New York, NY, USA: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. 1996. Corpus-Based Investigation of Language Use. *Annual Review of Applied Linguistics* 16: 114-136.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *The Longman Grammar of Spoken and Written English*. Longman.

Biber, D. and Reppen, R. 1998. Comparing native and learner perspectives on English grammar: a study of complement clauses. In S. Granger (ed.). *Learner English on Computer*, pp. 145-158. London: Addison Wesley Longman.

Bowker, L. 1998. Using Specialized Monolingual Native-Language Corpora As A Translation Resource: A Pilot Study. *Meta* XLIII(4).

Bryan, M. 1997. *SGML and HTML explained*. Harlow, England ; Reading, Mass.: Addison Wesley Longman.

Cabré, M. T. and Sager, J. C. 1999. *Terminology : theory, methods, and applications*. Amsterdam ; Philadelphia: J. Benjamins Pub. Co.

Cavaglià, G. and Kilgarriff, A. 2001. Corpora from the Web. *Proceedings of the 4th CLUK colloquium Sheffield*.

Chafe, W., DuBois, J. and Thompson., S. 1991. Towards A New Corpus Of Spoken American English. In K. Aijmer and B. Altenberg (eds.). *English Corpus Linguistics*, pp. 64–82. London: Longman.

Chen, H.–J. H. 1998. Underuse, overuse and misuse in Taiwanese EFL learner corpus. In S. Granger and J. Hung (eds.). *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 25–28. The Chinese University of Hong Kong.

Chi, A., Pui–yui, K. W. and Chau–ping, M. W. 1994. Collocational problems amongst ESL learners: a corpus–based study. In L. Flowerdew and A. K. K. Tong (eds.). *Entering Text*: The Hong Kong University of Science and Technology.

Chomsky, N. 1957. *Syntactic structures*. 's–Gravenhage: Mouton.

Church, K., Gale, W., Hanks, P. and Hindle, D. 1991. Using statistics in lexical analysis. *Lexical Acquisition: Exploiting On–Line Resources to Build a Lexicon*, pp. 115–164. Lawrence: Erlbaum.

Church, K. W. and Mercey, R. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19(1): 1–24.

Coates, J. 1983. *The Semantics of the Modern Auxiliaries*. London: Croom Helm.

Collins, P. 1991. The Modal of Obligation and Necessity in Australian English. In K. Aijmer

and B. Altenberg (eds.). *English Corpus Linguistics*, pp. 145–165. London: Longman.
Connor, U. and Kaplan, R. B. 1987. *Writing across languages : analysis of L b2 s text*. Reading, Mass.: Addison-Wesley.
Cover, R. 1992. Standard Generalized Markup Language ISO 8879:1986 Annotated Bibliography and List of Resources. Version 2.0
Cowie, A. P. 1992. Multiword Lexical Units and Communicative Language Teaching. In Arnaud and Bjoint (eds.). *Vocabulary and Applied Linguistics*, pp. 1–12. London: Macmillan.
Cruse, D. A. 1986. *Lexical semantics*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.
Crystal, D. 1991. *A dictionary of linguistics and phonetics*. Oxford, UK ; Cambridge, Mass., USA: B. Blackwell.
Dagneaux, E., Denness, S. and Granger, S. 1998. Computer-aided Error Analysis. *System* 26(2): 163–174.
Daille, B. 1995. Combined Approach for Terminology Extraction: Lexical Statistics *and Linguistic Filtering*. *UCREL Technical Papers* 5.
Diaz-Negrillo, A. and Fernandez-Dominguez, J. 2006. Error Tagging Systems for Learner Corpora. *RESLA* 19: 83–102.
Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1): 61–74.
Eco, U. 1976. *A theory of semiotics*. Bloomington: Indiana University Press.
Ellis, N. C. 2008. Phraseology: The periphery and the heart of language. In F. Meunier and S. Granger (eds.). *Phraseology in foreign language learning and teaching*, Amsterdam ; Philadelphia: John Benjamins.
Even-Zohar, I. 1990. Introduction to Polysystem Studies. *Poetics Today* 11(1): 1–6.

Fellbaum, C. 1998. *WordNet : an electronic lexical database*. Cambridge, Mass.: MIT Press.
Fillmore, C. 1992. Corpus Linuistics or Computer-aided armchair linguistics. In J. Svartvik (ed.). *Directions in corpus linguistics. Proceedings of Nobel Symposium 82. Stockholm 4-8 August 1991*, pp. 35-60. Berlin: Mouton.
Firth, J. R. 1957. A synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Blackwell.
Firth, J. R. 1968. Linguistics and Translation. In F. R. Palmer (ed.). *Selected Papers of J.R. Firth 1952-59*, pp. 84-95. London: Longman.
Fisiak, J. 1980. *Theoretical issues in contrastive linguistics*. Amsterdam: Benjamins.
Fisiak, J. 1990. *Further insights into contrastive analysis*. Amsterdam ; Philadelphia: J. Benjamins Pub. Co.
Fisiak, J. and Uniwersytet im. Adama Mickiewicza w Poznaniu. 1984. *Contrastive linguistics : prospects and problems*. Berlin ; New York: Mouton.
Fletcher, W. H. 2004. Making the Web More Useful as a Source for Linguistic Corpora. In U. Connor and T. A. Upton (eds.). *Applied corpus linguistics : a multidimensional perspective. Language and computers no. 52*, pp. 191-205. Amsterdam: Rodopi.
Flowerdew, L. 1998. Integrating expert and interlanguage computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes* 17(4): 329-345.
Fontenelle, T. 1994. Design and structure of the prototype collocation lexicon: Part *I: What are collocations?* - the DECIDE approach.
Fox, G. 1991. Context Dependency and Transitivity in English. In T. Johns and P. King (eds.). *Classroom Concordancing, ELR Journal 4*, pp. 17-25. University of Birmingham.

Francis, W. N., Ku*cera, H. and Mackie, A. W. 1982. *Frequency analysis of English usage : lexicon and grammar*. Boston: Houghton Mifflin.

Friedbichler, I. and Friedbichler, M. 1997. *The potential of domanin-specific target-language corpora for the translator's workbench*. *Corpus use and learning to translate*, Bertinoro.

Fries, C. C. 1940. *American English Grammar. Monograph 10*. New York: National Coucil of Teachers of English.

Fries, C. C. 1952. *The structure of English; an introduction to the construction of English sentences*. New York,: Harcourt, Brace.

Garnier, S., Tall, Y., Fissaha, S. and Haller, J. 2003. Learner Corpora: Design, Development and Applications - Development of NLP tools for CALL based on learner corpora (German as a foreign language). In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics 2003*, pp. 246-252. Lancaster University

Gaussier, E., Lange, J.-M. and Meunier, F. 1992. Towards Bilingual Terminology. *Proceedings of the Joint ALLC/ACH Conference*, pp. 121-124. Oxford: Oxford University Press.

Gaussier, E. and Lange., J.-M. 1994. Some Methods for the Extraction of Bilingual Terminology. In D. Jones (ed.). *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pp. 242-247. Manchester: UMIST.

Gavioli, L. 1996. Corpora and the concordancer in learning ESP: an experiment in a course of interpreters and translators. *the 18th Congress of the Associazione Italiana di Anglistica*, Genova.

Gavioli, L. and Zanettin, F. 1997. Comparable corpora and translation: a pedagogic perspective. *the first international conference on Corpus Use and Learning to Translate*, Bertinoro.

George, H. V. 1963. *Report on a Verb-Form Frequency Count* Hyderabad: Central Istitute of English.
Granger, S. 1983. The Be + Past Participle Construction. *Spoken English With Special Emphasis On The Passive*, Amsterdam: Elsevier Science Publishers.
Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg and M. Johansson (eds.). *Languages in Contrast. Text-based cross-linguistic studies. Lund Studies in English 88*, pp. 37-51. Lund: Lund University Press.
Granger, S. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (ed.). *Phraseology: theory, analysis and applications*, pp. 145-160. Oxford: Oxford University Press.
Granger, S. 2003. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In S. Granger, J. Lerot and S. Petch-Tyson (eds.). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, pp. 17-29. Amsterdam & Atlanta: Rodopi.
Granger, S. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* 20(3): 465-480.
Granger, S. 2003. The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* (special issue on corpus linguistics).
Granger, S. 2004. Computer learner corpus research: current status and future prospects. In U. Connor and T. A. Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective*, pp. 123-145. Amsterdam & Atlanta: Rodopi.

Granger, S., Hung, J. and Petch-Tyson, S. 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam ; Philadelphia: John Benjamins Pub.

Granger, S., Lerot, J. and Petch-Tyson, S. 2003. *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam ; New York: Rodopi.

Granger, S. and Tyson, S. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15: 19-29.

Granger, S. and Wynne, M. 1999. Optimising measures of lexical variation in EFL learner corpora. In J. M. Kirk (ed.). *Corpora Galore*, pp. 249-257. Amsterdam and Atlanta: Rodopi.

Guli, A. and Signorini, A. 2005. The indexable web is more than 11.5 billion pages. *WWW2005 - 14th World Wide Web Conference*.

Haan, P. d. 1989. *Postmodifying clauses in the English noun phrase : a corpus-based study*. Amsterdam Atlanta, GA: Rodopi.

Haan, P. d. 1991. On The Exploration Of Corpus Data By Means Of Problem-Oriented Tagging: Postmodifying Clauses In The English Noun Phrase. In Johansson and Sternstrom (eds.). *English Computer Corpora*, pp. 51-66. Berlin: Mouton de Gruyter.

Haas, W. 1968. The Theory of Translation. In G. H. R. Parkinson (ed.). *The Theory of Meaning*, pp. 86-108. Oxford: Oxford University Press.

Harris, Z. S. 1982. *A grammar of English on mathematical principles*. New York: Wiley.

Harris, Z. S. 1991. *A theory of language and information : a mathematical approach*. Oxford: Oxford University Press.

Heid, U. 2009. Corpus linguistics and lexicography. In A. Lüdeling and M. Kytö (eds.). *Corpus linguistics : an international handbook*

pp. 131–153. Berlin – New York: Walter de Gruyter.
Higgins, J. 1991. Fuel for learning: The neglected element of textbooks and CALL. *CAELL Journal* 2(2): 3–7.
Hill, L. A. 1960. The Sequence of Tenses with if-clauses. *Language Learning* 10(3): 165–178.
Hoey, M. and Arts and Humanities Research Board (Great Britain). 2005. *Lexical priming : a new theory of words and language*. London ; New York: Routledge/AHRB.
Hymes, D. 1981. On Communicative Teaching. In C. J. Brunflt and K. Johnson (eds.). *The Communicative Approach to Language Teaching*, pp. 7–16. London: Oxford University Press.
Ide, N., Reppen, R. and Suderman, K. 2002. The American National Corpus: More Than the Web Can Provide. *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, pp. 839–844. Las Palmas, Canary Islands, Spain.
Johansson, C. 1980. The LOB Corpus of British English Texts: Presentation and Comments. *ALLC Journal* 1: 25–36.
Johansson, C. 1993. Whose and of which with nonpersonal antecedents in written and spoken English. In C. Souter and E. Atwell (eds.). *Corpus-based computational linguistics*, Amsterdam: Rodopi.
Johansson, C. 1995. The Relativizers Whose and Of Which. *Present-Day Englsh: Description And Theory*, Uppsala: Uppsala University.
Johansson, C. and Norheim, E. H. 1988. The Subjunctive in British and American English. *ICAME Journal* 12: 27–36.
Johns, T. 1986. Microconcord: a language-learner's research tool. *System* 14(2): 151–162.
Johns, T. 1988. Whence and Whither Classroom Concordancing. In T. Bongaerts, P. de Haan, S. Lobbe, and H. Wekker (ed.). *Computer Applications in Language Learning*, pp. 9–27.

Johns, T. 1991. From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning. *ELR-Journal* 4: 27-45.

Johns, T. 1991. Should you be persuaded - two samples of data-driven learning materials. In T. Johns and P. King (eds.). *Classroom Concordancing. ELR Journal 4*, pp. 1-16. University of Birmingham: Centre for English Language Studies.

Joos, M. 1964. *The English verb: form and meanings*. Madison,: University of Wisconsin Press.

Jurafsky, D. and Martin, J. H. 2000. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall.

Justeson, J. S. and Katz, S. M. 1995. Technical Terminology: ome Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1: 9-27.

Kamakura, Y. 2007. Phraseology in a Learner Corpus Compared with the Phraseology of UK and US Students. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.). *Corpus Linguistics 2007*: University of Birmingham.

Kehoe, A. and Gee, M. 2007. New corpora from the web: making web text more 'text-like'. In I. T. Päivi Pahta, Terttu Nevalainen and Jukka Tyrkkö (ed.). *Studies in Variation, Contacts and Change in English 2: Towards Multimedia in Corpus Studies*.

Kennedy, G. 1991. Between and through: The company they keep and the functions they serve. In K. Aijmer, B. Altenberg and J. Svartvik (eds.). *English Corpus Linguistics: Studies in Honor of Jan Svartvirk*, London: Longman.

Kennedy, G. 1991. Preferred Ways Of Putting Things With Implications For Language Teaching.

Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm.
Kennedy, G. D. 1998. *An introduction to corpus linguistics*. London ; New York: Longman.
Kettemann, B. 1995. Concordancing in English language teaching. *TELL and CELL* 4(95): 4–15.
Kilgarriff, A. 2003. What computers can and cannot do for lexicography, or, Us precision, them recall. *Proceedings of ASIALEX*, Tokyo.
Kilgarriff, A. and Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3): 333–347.
Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. The sketch engine. *Proceedings of EURALEX*, Lorient.
Krenn, B. and Samuelsson, C. 1997. *The Linguist's Guide to Statistics*. University of the Saarland.
Lawrence, S., Giles, C. L. and Bollacker, K. 1999. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer* 32(6): 67–71–67–71.
Leńko–Szymańska, A. 2000. How to trace the growth in learner's active vocabulary. A corpus–based study. *The 4th International Conference on Teaching and Language Corpora*.
Leńko–Szymańska, A. 2000. Passive and active vocabulary knowledge in advanced learners of English. In B. Lewandowska–Tomaszczyk and J. P. Melia (eds.). *PALC' 99: Practical Applications in Language Corpora*, pp. 287–302. Frankfurt: Peter Lang.
Lorenz, G. 1997. *Introducing a Learner Corpus of English Writing: The Hidden Potentials of Adjective Intensification*. Amsterdam: Rodopi.
Louw, B. 1991. Classroom Concordancing of Delexical Forms and the Case for Integrating Language and Literature. In T. Johns and P. King (eds.). *Classroom concordancing, ELR Journal 4*, pp. 151–178. University of Birmingham.

Lozano, C. and Mendikoetxea, A. 2007. Learner Corpora and the Acquisition of Word Order: A Study of the Production of Verb-Subject Structures in L2 English. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.). *Linguistics 2007*: University of Birmingham.

Mallikamas, P. 1999. Applications of Corpora in Language Teaching. *Thai TESOL Bulletin* 12(1): 1-17.

Manning, C. D. and Schütze, H. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Mark, K. 1998. A parallel learner corpus approach to English curriculum development at a Japanese university. In S. Granger and J. Hung (eds.). *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 89-90. The Chinese University of Hong Kong.

Mark, K. 1998. The Significance of Learner Corpus Data in Relation to the Problems of Language Teaching. *Bulletin of General Education*, pp. 77-90. Meiji University.

Meyer, C. F. 1992. *Apposition in contemporary english*. Cambridge [England] ; New York: Cambridge University Press.

Meyer, C. F., Grabowski, R., Han, H.-Y., Mantzouranis, K. and Moses, S. 2003. The World Wide Web as Linguistic Corpus. *Language and Computer* 46: 241-254.

Milton, J. 1998 WORDPILOT: enabling learners to navigate lexical universes. . In S. Granger and J. Hung (eds.). *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, : The Chinese University of Hong Kong

Milton, J. and Tsang, E. 1993. A corpus-based study of logical connectors in EFL students' writing. In R. Pemberton and E. Tsang (eds.). *Studies in Lexis*, pp. 215-246. Language Centre,

The Hong Kong University of Science and Technology.
Mindt, D. and Weber, C. 1989. Prepositions in American and British English. *World Englishes* 8: 229–238.
Mishan, F. 2004. Authenticating corpora for language learning: a problem and its solution. *ELT Journal* 58(3): 219–227.
Muehileisen, V. 2006. Introducing the SILS Learners' Corpus: A Tool for Writing Curriculum Development. *Waseda Global Forum* 3: 119–125.
Murison-Bowie, S. 1996. Linguistic corpora and language teaching. *Annual Review of Applied Linguistics* 16: 182–199.
Naggaum, E. 1993. COMP.TEXT.SGML – Frequently Asked Questions. Version 1.2.
Nakamura, S. 2007. Comparison of Features of Texts Translated by Professional and Learner Translators. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.). *Corpus Linguistics 2007*: University of Birmingham.
Neff, J., Ballesteros, F., Dafouz, E., Martínez, F. and Rica, J.-P. 2003. Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics 2003*, pp. 562–571. Lancaster University
Neff, J., Dafouz, E., Herrera, H., Martínez, F., Rica, J. P., Díez, M., Prieto, R. and Sancho, C. 2003. Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance. In S. Granger, J. Lerot and S. Petch-Tyson (eds.). *Extending the scope of corpus-based research New applications, new challenges*, pp. 211–230. Amsterdam & Atlanta: Rodopi.
Nehls, D. 1979. *Studies in contrastive linguistics and error analysis*. Heidelberg: Groos.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam ; Philadelphia: J. Benjamins Pub. Co.

Newman, A. 1980. *Mapping translation equivalence*. Leuven, Belgium: Acco.

Nicholls, D. 2003. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics 2003*, pp. 572–581. Lancaster University.

Oakes, M. P. and NetLibrary Inc. 1998. Statistics for corpus linguistics. *Edinburgh textbooks in empirical linguistics*, pp. xvi, 287 p. Edinburgh: Edinburgh University Press.

Oketch, V. A. 2007. A Multilingual Learner Corpus: A Description of the Kenyan Multilingual Individual and His L4 Interlanguage. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.). *Corpus Linguistics 2007*: University of Birmingham.

Ota, A. 1963. *Tense and Aspects of Present-Day American English*. Tokyo: Kenkyusha.

Pan, W. g. and Mun, T. W. 2007. *Contrastive linguistics : history, philosophy and methodology*. London ; New York: Continuum.

Paquot, M. 2005. EAP vocabulary in learner corpora: a cross-linguistic perspective. In C. Cosme, C. Gouverneur, F. Meunier and M. Paquot (eds.). *Proceedings of the Phraseology 2005 Conference*, pp. 323–326. Louvain-la-Neuve.

Pendar, N. and Chapelle, C. A. 2008. Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal* 25(2): 189–206.

Pravec, N. A. 2002. Survey of learner corpora. *ICAME Journal* 26: 81–114.

Quirk, R. 1985. *A Comprehensive grammar of the English language*. London ; New York: Longman.

Reder, S., Harris, K. and Setzler, K. 2003. The Multimedia Adult Learner Corpus. *TESOL Quarterly* 37: 546–557.

Renouf, A. 1993. A Word in Time: First Findings from the Investigation of Dynamc Text. In J. Aarts, P. d. Haan and N. Oostdijk (eds.). *English Language Corpora: Design, Analysis and Exploitation*, pp. 277–288. Amsterdam: Rodopi.

Renouf, A. a. J. S. 1991. Collocation Frameworks in English. In K. Aijmer and B. Altenberg (eds.). *English Corpus Linguistics*, pp. 128–143. London: Longman.

Roumen, R. and Ster, T. V. D. 1993. CONTEXT: A New Concept in Computer-Aided Translation. *TKE'93: Terminology and Knowledge Engineering*, pp. 215–220. Frankfurt: Indeks Verlag.

Rundell, M. 1996. The Corpus of the Future, and the Future of the Corpus. *New trends in Reference Science*.

Satoh, K. and Fang, C. A. 1998. A corpus-based study of the grammar and lexis of Japanese learners' English. In S. Granger and J. Hung (eds.). *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 103–104. The Chinese University of Hong Kong.

Saussure, F. d. 1959. *Course in general linguistics*. New York,: Philosophical Library.

Schmied, J. and Schaffler, H. 1996. Approaching Translationese Through Parallel and Translation. *Synchronic corpus linguistics Pepers from the sixteenth International Conference on English Language Research on Computerized Corpora*, Toronto Rodopi.

Shamaa, N. 1978. A Linguistic Analysis of Some Problems of Arabic to English Translation. University of Oxford.

Sinclair, J. 1986. Basic Computer Processing Of Long Texts. In G. Leech and C. N. Candlin (eds.). *Computers in English Language Teaching and Research*, pp. 185–203. Harlow, Essex: Longman.

Sinclair, J. 1987. The Nature of Evidence. In J. Sinclair (ed.). *Looking Up: An Account of the COBUILD Project*, London: Collins.
Sinclair, J. (ed.) 1990. *Collins Cobuild English Grammar*. Collins.
Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford ; New York: Oxford University Press.
Sinclair, J. 1992. Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik (ed.). *Directions in corpus linguistics : proceedings of Nobel Symposium 82, Stokholm, 4-8 August 1991*, pp. 35-60. Berlin ; New York: Mouton de Gruyter.
Sinclair, J. 1998. Corpus Evidence In Language Description. In A. Wichmann, S. Filegelstone, A. McEnery and G. Knowles (eds.). *Teaching and language corpora*, Harlow: Longman.
Smith, J. M. 1992. *SGML And Related Standards: Document Description And Processing Languages*. New York: Ellis Horwood.
Souter, C. and Atwell, E. 1993. *Corpus-based computational linguistics*. Amsterdam: Rodopi.
Sperberg-McQueen, C. M. and Burnard, L. 1990. Guidelines for Electronic Text Encoding and Interchange. Chapter 42. *Text Encoding Initiative*, Chicago.
Sterkenburg, P. G. J. v. 2003. *A practical guide to lexicography / edited by Piet van Sterkenburg*. Amsterdam ; Philadelphia: John Benjamins Pub.
Stevens, V. 1995. Concordancing With Language Learners: Why? When? What? . *CAELL Journal* 6(2): 2-10.
Sugiura, M. 2002. Collocational Knowledge of L2 Learners of English: A Case Study of Japanese Learners. In T. Saito, J. Nakamura and S. Yamazaki (eds.). *English Corpus Linguistics in Japan*, pp. 303-323. Amsterdam & Atlanta: Lodopi.

Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.
Summers, D. 2005. Corpus Lexicography – The importance of representativeness in relation to frequency.
Svartvik, J. 1966. *On voice in the English verb*. The Hague,: Mouton.
Svartvik, J. 1992. Lexis in English Language Corpora. In H. Tommola and K. Varantola (eds.). *Euralex 92. Proceeding I and II*, pp. 17–31. Tampere: Tampereen Yliopisto.
Svartvik, J. 2007. Corpus linguistics 25+ years on. In R. Facchinetti (ed.). *Corpus linguistics 25 years on*, pp. 11–26. Amsterdam: Rodopi.
Tagnin, S. 2003. A multilingual learner corpus in Brazil. In P. R. Dawn Archer, Andrew Wilson and Tony McEnery (ed.). *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster University.
Tan, M. 2005. Authentic language or language errors? Lessons from a learner corpus. *ELT Journal* 59(2): 126–134.
Thorndike, E. I. 1921. *Teacher's Workbook*. New York: Columbia Teachers College.
Thorndike, E. I. and Lorge, I. 1944. *The Teacher's Workbook of 30,000 Words*. New York: Columbia University Press.
Tollenaere, E. d. 1973. The Problem of the Context in Computer-Aided Lexicography. In A. J. Aitken, R. W. Bailey and N. Hamilton-Smith. (eds.). *The Computer and Literary Studies*, Edinburgh: Edinburgh University Press.
Tono, Y. 2003. Learner corpora: design, development and applications. In P. R. D. Archer, A. Wilson and T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 conference*, pp. 800–809. Lancaster University.
Tono, Y. and Aoki, M. 1998. Developing the optimal learning list of irregular verbs based on the

native and learner corpora. In S. Granger and J. Hung (eds.). *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 113–118. The Chinese University of Hong Kong.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T. and Izumi, E. 2001. The Standard Speaking Test (SST) Corpus: A 1 million–word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In S. Lee (ed.). *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary. The Second Asialex International Congress*, pp. 257–262. Yonsei University, Korea.

Tottie, G. 1991. *Negation in English speech and writing : a study in variation*. San Diego: Academic Press.

Toury, G. 1978. The Nature and Role of Norms in Literary Translation. In J. S. Holmes, J. Lambert and R. v. d. Broeck (eds.). *Literature and Translation*, pp. 83–100. Leuven: ACCO.

Toury, G. 1980. *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics.

Toury, G. 1985. A Rationale for Descriptive Translation Studies. In T. Hermans (ed.). *The manipulation of Literature: Studies in Literary Translation*, pp. 16–41. Sydney: Croom Helm.

Tribble, C. 1989. The Use of Text Structuring Vocabulary in Native and Non–Native *Speaker Writing*. *MUESLI News* June: 11–13.

Tribble, C. 1997. Improvising Corpora For ELT: Quick–And–Dirty Ways Of Developing *Corpora For Language Teaching*. In M. J and B. Lewandowska–Tomaszczyk (eds.). *PALC '97 Proceedings*, Lodz: Lodz University Press.

Tribble, C. 2000. Genres, Keywords, Teaching: Towards A Pedagogic Account Of The *Language Of Project Proposals* In L. Burnard and T. McEnery (eds.). *Rethinking language pedagogy from a*

corpus perspective: papers from the third international conference on teaching and language corpora, Hamburg: Peter Lang.
Tribble, C. and Jones, G. 1990. *Concordances in the Classroom*. Longman.
Van Herwijnen, E. 1990. *Practical SGML*. Dordrecht ; Boston: Kluwer Academic Publishers.
Van Rooy, B. and Schäfer, L. 2003. Automatic POS tagging of a learner corpus: The influence of learner error on tagger accuracy. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 conference*, pp. 835–844. Lancaster University.
Vanderauwera, R. 1985. *Dutch Novels Translated into English: The Transformation of a Minority Literature.* Amsterdam: Rodopi.
Varantola, K. 1997. Translators, dictionaries and text corpora. *Corpus use and learning to translate Bertinoro*.
Vermeer, H. J. 1983. *Aufs?tze zur Translation theories*. Heidelberg: Groos.
Volk, M. 2002. Using the Web as Corpus for Linguistic Research. In R. Pajusalu and T. Hennoste (eds.). *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim.*, Department of General Linguistics: University of Tartu.
Waibel, B. 2007. *Phrasal verbs in learner English: a corpus-based study of German and Italian students*. Freiburg: Albert-Ludwigs-Universität Freiburg.
Wang, S. 1991. A Corpus Study of English Conditionals. Victoria University of Wellington.
West, M. 1953. *A General Service List of English Words*. London: Longman.
Wible, D. and Huang, P.-Y. 2003. Using learner corpora to examine L2 acquisition of tense-aspect markings. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Corpus Linguistics 2003*, pp. 889–898. Lancaster University.

Widdowson, H. G. 1980. Models and fictions. *Applied Linguistics* 1(2): 165–170.
Widdowson, H. G. 1983. *Learning purpose and language use*. Oxford ; New York: Oxford University Press.
Widdowson, H. G. 1984. *Explorations in Applied Linguistics 2*. Oxford: Oxford University Press.
Widdowson, H. G. 1990. *Aspects of language teaching*. Oxford ; New York: Oxford University Press.
Widdowson, H. G. 1991. The description and prescription of language. In J. E. Alatis (ed.). *Georgetown University round table on language and linguistics 1991*, pp. 11–24. Washington D.C.: Georgetown University Press.
Wynne, M. 2005. Stylistics: Corpus Approaches. In E. K. Brown and A. Anderson (eds.). *Encyclopedia of language & linguistics*, pp. 223–226. Amsterdam ; Boston: Elsevier.
Zanettin, F. 1998. Bilingual Comparable Corpora and the Training of Translators. *Meta* 43(4).

ประวัติผู้เขียน

วิโรจน์ อรุณมานะกุล เป็นอาจารย์ประจำภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ จากจุฬาลงกรณ์มหาวิทยาลัย ปริญญาโทอักษรศาสตรมหาบัณฑิต สาขาภาษาศาสตร์ จากจุฬาลงกรณ์มหาวิทยาลัย และปริญญาเอก สาขาภาษาศาสตร์คอมพิวเตอร์ จากมหาวิทยาลัยจอร์จทาวน์ ประเทศสหรัฐอเมริกา สมรสกับ อาจารย์วิไลวรรณ อรุณมานะกุล อาจารย์ประจำสถาบันภาษา มหาวิทยาลัยธรรมศาสตร์ ไม่มีบุตร มีแต่สัตว์เลี้ยงเป็นสุนัขและแมวจำนวนหนึ่ง

Home page : http://pioneer.chula.ac.th/~awirote/

หนังสือภาษาศาสตร์คลังข้อมูลเล่มนี้ ชี้ให้เห็นถึงบทบาทและความสำคัญของคลังข้อมูลภาษาในงานด้านต่างๆ ได้แก่ การวิจัยทางภาษาศาสตร์ การเรียนการสอนภาษา การแปลภาษา และการประมวลผลภาษา หนังสือเล่มนี้ได้ให้ความรู้เบื้องต้นเกี่ยวกับการรวบรวมข้อมูลภาษาเพื่อจัดทำเป็นคลังข้อมูลภาษาสำหรับใช้งาน

วิโรจน์ อรุณมานะกุล เป็นอาจารย์ประจำภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย มีผลงานและรับผิดชอบการสอนในด้านภาษาศาสตร์คลังข้อมูล และภาษาศาสตร์คอมพิวเตอร์